Exploring global and local patterns in the correlation of geographic distances and morphosyntactic variation in Swiss German

Jeszenszky, Péter ; Stoeckle, Philipp ; Glaser, Elvira ; Weibel, Robert

Abstract: Using data from a Swiss German dialect syntax survey, this study aims to explore, in a spatially differentiated manner, the correlation between dialectal variation and geographic distances. A linguistic distance was expressed by a measure aggregated from 60 survey questions. To operationalize the possibility of language contact, Euclidean distance, as well as travel times in 2000, 1950 and 1850 between survey sites were used. Going beyond previous work by others, we also explore the covariation of geographic and linguistic distances at the local level, focusing on spatial subsets and individual survey sites, thus being able to paint a more differentiated picture. With the diverse physical landscape of Switzerland making an impact on potential language contact, we find that travel times are a better predictor than Euclidean distance for the syntactic variation in Swiss German dialects. However, on the local scale the difference is not always significant, depending on prevalent topography.

DOI: https://doi.org/10.1017/jlg.2017.5

EXPLORING GLOBAL AND LOCAL PATTERNS IN THE CORRELATION OF GEOGRAPHIC DISTANCES AND MORPHOSYNTACTIC VARIATION IN SWISS GERMAN

Péter Jeszenszky[1], Philipp Stöckle[2], Elvira Glaser[2], Robert Weibel[1]

[1] Department of Geography, University of Zurich (UZH),
Winterthurerstrasse 190, CH-8057, Zurich
Email: peter.jeszenszky@geo.uzh.ch
Tel.: +41-44-635-51-31

[2] German Department, University of Zurich (UZH),
Schönberggasse 2, CH-8001, Zurich

Short title:
Syntactic variation explained by geographic distances

Abstract

Using data of a Swiss German dialect syntax survey this study aims at exploring, in a spatially differentiated manner, the correlation between dialectal variation and geographic distances. A linguistic distance was expressed by a measure aggregated from 60 survey questions. To operationalise the possibility of language contact, Euclidean distance and travel times in 2000, 1950 and 1850 between survey sites were used. Going beyond previous work by others, we also explore the covariation of geographic and linguistic distances at the local scale, focusing on spatial subsets and individual survey sites, thus being able to paint a more differentiated picture. With the diverse physical landscape of Switzerland making an impact on potential language contact, we find that travel times are a better predictor than Euclidean distance for the syntactic variation in Swiss German dialects. However, on the local scale the difference is not always significant, depending on prevalent topography.

Acknowledgements

The research reported in this paper represents part of the PhD project of the first author. Funding by the Swiss National Science Foundation through project “Modelling morphosyntactic area formation in Swiss German (SynMod)” (SNF Project no. CR12I1_140716 & CR12I1_162760) is gratefully acknowledged. We are grateful to the Institute for Transport Planning and Systems at ETH Zurich for providing their travel time data, and to Gabi Bart and Sandro Bachmann of the Syntactic Atlas of German-speaking Switzerland (SADS) project for the provision of, and professional help with, the syntax data. Further, we would like to acknowledge the comments of Carlota de Benito Moreno and the anonymous reviewers.

1. INTRODUCTION

1.1 Motivation and Hypotheses

As geographic factors have the potential to crucially influence language contact and thus impact on linguistic variation (Wang & Cavalli-Sforza, 1986), there has been a stream of research in modern dialectology that has investigated the relationship of geographic distances with linguistic variation (e.g. Gooskens, 2004; Heeringa & Nerbonne, 2001; Nerbonne & Kleiweg, 2007; Pickl, Spettl, Pröll, Elspaß, König & Schmidt, 2014; Séguy, 1971; Shackleton, 2007; Szmrecsanyi, 2012). The strong correlation that was observed in most of such studies led Nerbonne and Kleiweg (2007:154) to formulate the 'Fundamental Dialectological Postulate' (FDP), which posits that "geographically proximate varieties tend to be more similar than distant ones". As this observation also holds more generally for many (or most) other geographically distributed phenomena, the discipline of geography knows a very similar postulate, first formulated by Tobler (1970:236): "Everything is related to everything else, but near things are more related than distant things". The universality of this observation later led others to call it "Tobler's First Law of Geography" (Sui, 2004).

In essence, both postulates describe an effect that is commonly known as spatial autocorrelation (Griffith, 1987), which describes the degree of spatial dependency in a spatially distributed variable. In dialectology it is the linguistic similarity of spatially distributed language varieties. However, even though spatial autocorrelation is manifested in many geographical variables, it is by no means guaranteed to be strongly present in all cases. Thus, for instance, Szmrecsanyi (2012) in his corpus-based study of English dialects found only a very weak effect, causing him to conclude that "geography is overrated".

This paper reports on a study that explores the correlation between the variation in Swiss German syntax and geographic distances. However morphosyntactic dialect variation has recently witnessed an upheaval in interest after a long time of neglect, correlation analysis of syntax against geographic distances has so far only been carried out by Spruit (2006, 2008) and Szmrecsanyi (2012, 2014) on Dutch and English dialects, respectively. The syntax dataset we used originates from the Syntactic Atlas of German-speaking Switzerland (SADS). Owing to the peculiarities of the SADS data — most importantly the fact that it features contribution from multiple respondents in various numbers per survey site — we use a particular measure to represent linguistic distance, similar to Speelman, Grondelaers and Geeraerts (2003). We adopt an aggregate variation approach (e.g. Nerbonne, 2009), aggregating 60 specific syntactic phenomena (variables), as represented by 60 questions in the SADS, to build a linguistic distance measure. Switzerland features a diverse surface topography, which in mountainous areas places constraints on transportation and communication routes, and thus imposes barriers to potential language contact between speakers. We therefore use different geographic distance measures to operationalise the possibility of language contact and to calculate correlations with the linguistic distance. Besides Euclidean distance (distance "as the crow flies") travel times of different points in time (1850, 1950, 2000) are also included. Finally, while previous studies by other authors were restricted to computing the correlation between linguistic variation and spatial distance at the *global* level only, we also study the correlation at *local* levels of smaller geographic areas, thus exploring local effects such as topographic barriers and, conversely, interconnectedness.

Besides physical barriers, other, (socio)cultural factors are considered to have influenced the evolution of Swiss German dialects (Hotzenköcherle, 1986), such as administrative subdivision and isolation in the times before the modern Swiss Confederation was formed, or religious borders (Roman catholic vs. protestant). Our interest is thus in finding out the degree to which geographic distance may explain linguistic variation. Using different times for the representation of travel times is motivated by the fact that convergence effects are noticeable in the evolution of Swiss German varieties (Christen, 1998). Also, syntax is assumed to differentiate at a slower rate over time than other linguistic levels (Longobardi & Guardiano, 2009). Thus, we expect the correlation to be best for the older dates of travel times.

Our work departs from the following hypotheses:

H1: Geographic distance is responsible for, and thus explains, the majority of the variance ($R^2 > 0.5$) found in Swiss German syntax, as represented in the SADS data.

H2: Among the geographic distance measures, travel time measures better reflect syntactic spatial variation than Euclidean distance.

H3: Older travel times better represent syntactic spatial variation.

1.2 Correlating Dialect Distances to Geographic Distances

Over the past decades, there have been a number of studies investigating the relation between linguistic variation and geographic distance. Séguy (1971) was the first to plot linguistic differences (expressed as counts of differences in lexical features between survey sites) as a function of geographic distance. He was also the first to identify a logarithmic correlation of linguistic difference with Euclidean distance. Wang and Cavalli-Sforza (1986) used Euclidean distance between Micronesian islands (as steps in a chain of islands) to quantify the spread of cognates. Heeringa and Nerbonne (2001) examined the degrees of association between geographical (Euclidean) and pronunciational differences in Dutch dialects using the Levenshtein distance. Spruit (2006) used an additive measure of differences to calculate a Hamming distance on the syntactic level, which he correlated with Euclidean distance. Spruit, Heeringa and Nerbonne (2009) investigated the influence of geography (expressed by Euclidean distance) for aggregate pronunciational, lexical and syntactic differences in Dutch. For the lexical and syntactic level Goebl's (1982) Weighted Identity Value was used while for pronunciation, the Levenshtein distance.

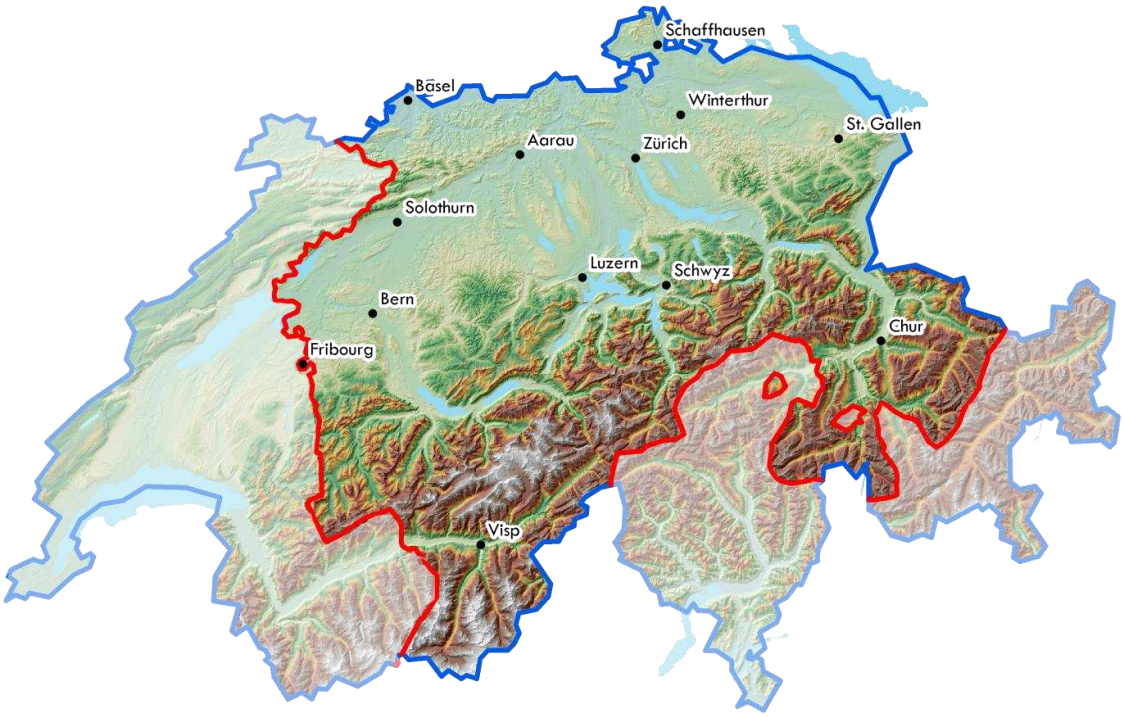
While most authors have used Euclidean distance as a measure of geographic distance, some chose other spatial quantifications to represent potential language contact with an increased degree of realism. Gooskens (2004) was the first to compare Euclidean distances to modern and old travel times to establish which one correlates better with Levenshtein and perceptual distances of Norwegian dialects. Euclidean distance and modern travel times produced the same correlation, while older travel times correlated better with both the Levenshtein and the perceptual distances. Haynie (2012) established the utility of cost distance modelling for studying historical language contact networks of Miwok languages in California, correlating it to a metric of recurrent sound correspondence.

As mentioned above, the seeming universality of the correlation of increasing linguistic differences with increasing geographic distance is expressed in the Fundamental Dialectological Postulate (FDP; Nerbonne and Kleiweg, 2007). Several authors subsequently have tested this hypothesis, with different results. While Shackleton (2007) investigated the correlation of phonetic distance to Euclidean distance for the traditional English dialects, reporting an R^2 of up to 0.77 (i.e. explaining 77% of linguistic variance) for a regression model accounting for regional differences, Szmrecsanyi (2012) found much lower values, using corpus-based data on morphosyntax of English dialects. In Szmrecsanyi's results, Euclidean distance explains a mere 4% of morphosyntactic variance; travel times fare only slightly better at 8%; Trudgill's (1974) Linguistic Gravity explains 24 %; and finally the maximum value of 32.5% is reached after clustering dialects into dialect groupings. Szmrecsanyi (2012:226) concludes that "It is fair to say that the FDP has failed this test".

More recently in linguistic research, the attention has shifted increasingly to syntax, owing to the increasing availability of this type of data in dialect atlases and inspired by observations that syntactic variation may also be geographically patterned (e.g. noted by Szmrecsanyi, 2014). In this paper, we focus also on morphosyntax, studied using the example of Swiss German dialects.

From a geographical perspective all of the above studies suffer from the crucial drawback of restricting the analysis to the — geographically speaking — global level, computing correlations for entire study areas, rather than exploring linguistic variation in more detail at the local level. Hence, they miss out on discovering regional differences in correlation structures, and on delivering possible

explanations of regionally different linguistic variation patterns. With additional local analyses, it may be possible to explain why high degrees of correlation had been reported in some studies, and low correlations in others.



Map 1 The diverse topography of Switzerland with some of selected cities in the German speaking area (within the red border).

2. DATA

2.1. Dialect Survey Data

About 65% of the population in Switzerland uses the German language. This study uses data gathered through the surveys of the Syntactic Atlas of German-speaking Switzerland (SADS; Bucheli & Glaser, 2002) in the area where German is the dominant language (Map 1). The SADS project was initiated in the year 2000 to map and study syntactical phenomena of Swiss German dialects. Most of the questions in the survey were either translation questions, where the respondents had to translate a Standard German sentence into their dialect, or multiple choice (MC) questions where respondents had to choose among given variants the ones they considered possible to use in their local dialect. In such cases respondents were also asked to pick a variant which they personally preferred over the others.

Close to 3,200 respondents participated in the surveys (see Bucheli & Glaser, 2002 for the selection criteria of respondents), providing answers to 118 written questions, corresponding to 50 morphosyntactic *phenomena*, with a tendency of choosing phenomena that were assumed to show spatial variation patterns. 383 places were surveyed, corresponding to about 25% of all German speaking municipalities in Switzerland, providing a rather dense sample compared to other dialect surveys. A key feature of the SADS is that multiple respondents occur per survey site, ranging between 3 and 26, with a median value of 6 to 7 respondents per site. The spatial distribution of the survey sites corresponds to the population density in Switzerland, as there are more survey sites in the densely populated Swiss Plateau (*"Mittelland"*) than in the sparsely populated mountainous regions.

Having multiple respondents per survey site provides the chance to better capture the endemic linguistic diversity that might be present within a place owing to age, social and professional differences, while capturing within-speaker variation is made possible through MC questions. This variation captured creates a very rich database for linguistic analyses, allowing to paint a highly differentiated picture of morphosyntax in Swiss German dialects, and enabling different perspectives on morphosyntactic variation (Stoeckle, 2016). However, the database also invariably represents more uncertainty than traditional atlas databases, which usually rely on a single respondent per survey site, providing a single variant per linguistic phenomenon. In order to increase the comparability with other studies, we thus worked with the preferred variants in the case of MC questions. In this study, 60 *variables* were used, with each variable corresponding to one aspect of one survey question (note that a survey question may cover more than one variable). This subset, which was selected in collaboration with the SADS group, covers the majority of the 50 morphosyntactic phenomena investigated in the SADS. The variables used in the study are listed in Table 1.

Table 1. The variables of the SADS used in the study

Question	Sentence in Standard German	Sentence in English	Short description of phenomenon
I1	Entschuldigung, ich habe zu wenig Kleingeld, um ein Billett zu lösen.	Excuse me, I don't have enough change in order to buy a ticket.	infinitival purposive clause: linkage (1)
I11	Aber jetzt habe ich mich gerade hingesetzt, um ein Buch zu lesen.	But I just sat down in order to read a book.	infinitival purposive clause: linkage (2)
I12	Fischstäbchen muss man doch gefroren anbraten.	Actually, fish fingers should be fried while still frozen.	copredicative participle
I13	Da wird gearbeitet	People are working here	expletive 'it' (impersonal passive)
I18	Soll ich welche kaufen?	Should I buy some?	partitive object (pronoun)
I19	Ich habe keine Ahnung, ob sie das Auto schon bezahlt hat.	I have no idea whether she has already paid for the car.	perfect auxiliary ('have') in subordinate clauses: position
I2	Wem will er denn die schönen Blumen bringen?	To whom does he want to bring those beautiful flowers?	prepositional dative marking (PDM) (1)
I20	Aber ich habe doch das Buch dir geschenkt.	But I gave the book as a present to you.	prepositional dative marking (PDM) (2)
I3	Oh, ich habe den Fritz kommen hören.	Oh, I heard Fritz coming.	perfect with 'hear': form and position of non-finite verb (IPP) (1)
I5	Der Korb ist umgekippt.	The basket is toppled over.	resultative: subject agreement (1)
I6	Wissen Sie, jetzt brauche ich sogar Tabletten zum einschlafen.	You know, now I even need pills in order to fall asleep.	infinitival purposive clause: linkage (3)
I7	Nein, das gehört meiner Schwester.	No, it belongs to my sister.	prepositional dative marking (PDM) (3)
I8	Aber ich habe im Fall schon gestern geholfen abzuwaschen.	But I already helped doing the dishes yesterday.	perfect with 'help': form and position of non-finite verb (IPP) (2)

I9	Also ich weiss auch nicht, ob er einmal heiraten will.	Well, I don't know if he ever wants to get married.	modal verb in subordinate clauses: position
II1	Hast du die Uhr flicken lassen?	Have you had the clock fixed?	infinitive particle (doubling/position) 'let' (1)
II11	Er hat die Hand immer noch eingebunden.	He has his arm still bandaged.	resultative: object agreement (2)
II13	Du musst die Milch aber heiss trinken!	But you have to drink the milk hot!	copredicative adjective
II18	Das ist der Mann, dem ich gestern den Weg gezeigt habe.	That's the man to whom I gave directions yesterday.	relative clause linkage: IO (1)
II19	Und dann ist ein Fuchs geschlichen gekommen!	And then a fox came creeping around!	verbal construction 'come' + motion verb
II2	Das ist doch die Frau, der ich schon lange das Buch bringen sollte.	This is the woman to whom I should have brought back the book long ago.	relative clause linkage: IO (2)
II20	Ich möchte aber ein Auto, das ich auch bezahlen kann!	But I want a car that I can actually pay for!	relative clause linkage: DO (3)
II22	Nein, das ist Peters [Dreirad].	No, that's Peter's. [tricycle]	predicative possessive (1)
II23	Nein, das ist Sandras [Dreirad].	No, that's Sandra's. [tricycle]	predicative possessive (2)
II28	Das ist der Mann, mit dem ich immer schwätze.	That's the man I always chat with.	relative clause linkage: PP (4)
II3	Er lässt den Schreiner kommen.	He is going to call the carpenter.	infinitive particle (doubling/position) 'let' (2)
II30	Der Hund des Lehrers	The teacher's dog	adnominal possessive
II32	Ich habe Fritz gesehen	I have seen Fritz.	personal name: definite article and case inflection (1)
II4	Du hast sicher viel zu erzählen!	You must have a lot to tell.	non-finite form with 'have to' (gerund)
II5	Ihr dürft alles liegen lassen.	You can leave everything.	infinitive particle (doubling/position) 'let' (3)
II7	Ich habe erst mit vierzig fahren gelernt.	I have only learnt to drive at forty.	perfect with 'learn': form and position of non-finite verb (IPP) (3)
II9	Nein, sie ist gerade verkauft worden.	No, it has just been sold.	passive auxiliary and agreement
III1	Wenn es so warm bleibt, fängt das Eis an zu schmelzen!	If it stays this warm, the ice will start to melt.	infinitive particle (position/doubling) 'begin' (1)
III10	Wenn sie dich erwischen, bekommst du den Fahrausweis entzogen!	If they catch you, your driver's license will be taken away.	'get'-passive
III11	Also mich erwischt keiner!	Well, no one will catch me!	personal pronoun (1sg): DO (1)

III12	Nimm die Suppe sofort weg, wenn sie zu kochen anfängt!	Take the soup off immediately, once it starts boiling.	infinitive particle (position/doubling) 'begin' (2)
III13	Er gibt sich einfach keine Mühe.	He just doesn't put any effort into it.	reflexive pronoun (3sgm) (1)
III16	Die Strasse ist schon seit einem Jahr aufgerissen.	The street has already been torn up for a year.	resultative: subject agreement (3)
III17	Wir müssen uns das überlegen.	We have to think about it.	reflexive pronoun (1pl) (2)
III2	Wen suchst du?	Who are you looking for?	interrogative pronoun: case (1)
III20	Er schaut nur für sich selbst.	He only thinks about himself.	reflexive pronoun (PP) (3)
III22	Sie ist grösser als ich.	She is taller than me.	comparative clause linkage (1)
III23	Hinkend ist er gelaufen.	He went home limping.	converb
III25	Sie gehen halt lieber schwimmen als laufen.	They would rather go for a swim than for a walk.	comparative clause linkage (2)
III28	Dann ist er ja älter, als ich gemeint habe.	So he is older than I expected.	comparative clause linkage (3)
III3	Für wen sind denn die Blumen?	Who are the flowers for?	interrogative pronoun: case (2)
III4	Die sind nicht für dich!	They are not for you!	personal pronoun (2sg): PP (2)
III5	Ich habe schon angefangen zu kochen.	I have already started cooking.	infinitive particle (position/doubling) 'begin' (3)
III7	Sie hat es mir gestern erzählt.	She told me yesterday [about expecting a baby].	personal pronouns: position (1)
III8	Sie findet es nicht gut, dass ich angefangen habe zu rauchen.	She doesn't find it good that I started smoking.	infinitive particle (position/doubling) 'begin' (4)
IV11	Doch, das ist im Fall er gewesen.	Yes, that must have been him!	personal pronoun (3sgm): subject (3)
IV14	Du musst das Licht anzünden, um zu lesen.	You have to turn the light on in order to read.	infinitival purposive clause: linkage (4)
IV17	Doch, das ist er sicher gewesen!	Yes, that was him for sure!	personal pronoun (3sgm): subject (4)
IV19	Ja, ich habe etwas ganz Schönes gekauft!	Yes, I have bought something really nice!	indefinite pronoun: position/doubling
IV21	Ich habe nicht gewusst, dass er so spät fahren gelernt hat.	I didn't know that he learnt to drive only so late.	perfect with 'learn': form and position of non-finite verb (IPP) (4)
IV25	Das glaubst du ja selber nicht, dass sie so früh lesen gelernt hat.	No way she learnt to read so young!	perfect with 'learn': form and position of non-finite verb (IPP) (5)
IV28	Ich habe es Fritz gegeben.	I gave it to Fritz.	personal name: definite article and case inflection (2)

IV3	Ich habe es ihm schon geschickt.	I have already sent it to him.	personal pronouns: position (2)
IV31	Das gefallen täte mir auch!	I would like it, too!	subjunctive auxiliary 'do' (position)
IV4	Wer ist das gewesen?	Who was it?	interrogative pronoun: case (3)
IV7	Jetzt kannst du anfangen.	Now you can start.	non-finite 'begin' (5) with modal verb

2.2. Travel Times

As visible on Map 1 Switzerland features a diverse surface topography that imposes constraints on transport routes and thus on possible contact paths between speakers (though recent construction of tunnels have generally increased accessibility). To account for these constraints, geographic distances were not only represented by Euclidean distance, but also by travel times. The travel times matrices by car and by public transportation were provided by the Institute for Transport Planning and Systems at ETH Zurich (Fröhlich, Frey, Reubi & Schiedt, 2004).

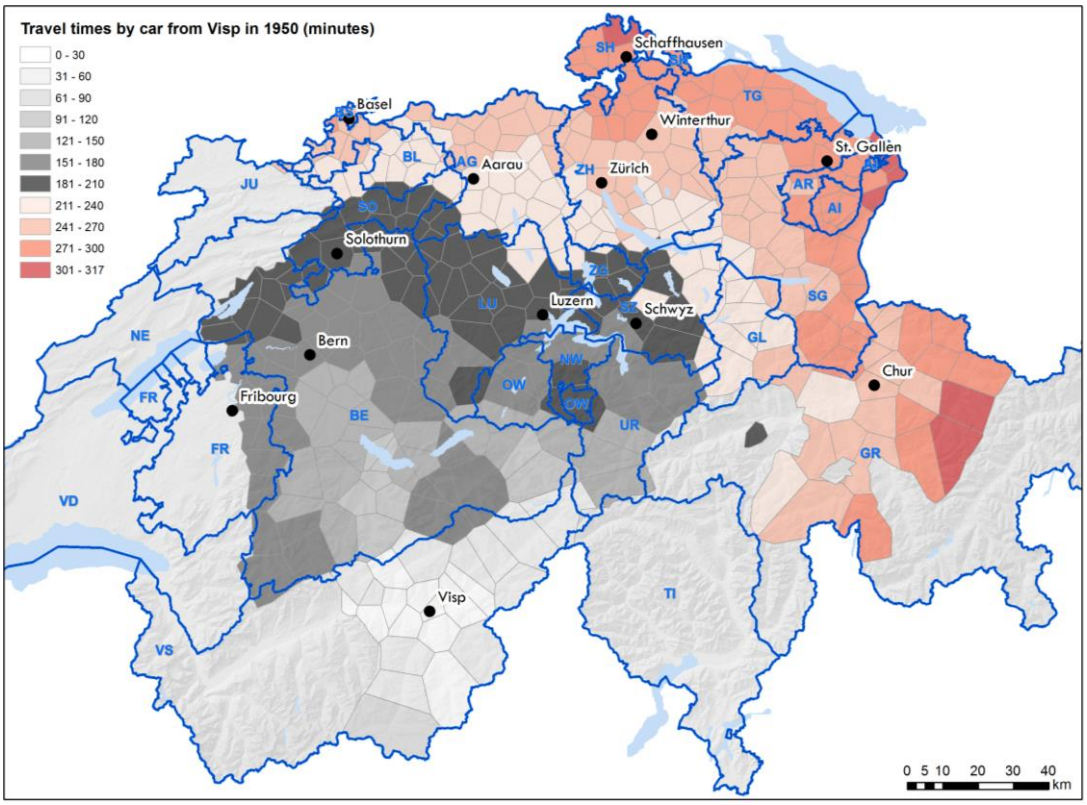
For this study travel times of 1850, 1950 and 2000 were used. For 1950 and 2000, car travel time matrices covered all 383 survey sites of the SADS. For 1850, the data was available only for 120 places, or roughly a third of the survey sites. However, as these places are equivalent to the district ("Bezirk") capitals, the spatial distribution generally follows that of the entire set of SADS points.

Each of the three time points is representative of a particular point in the development of the Swiss transportation infrastructure. 1850 is representative of the transportation network before railroads were built in Switzerland (the only railroad line that opened earlier than 1850 measured a mere 20 km in 1847). 1950 is representative of a road and train network before motorways, new fast train connections and commuter train systems were built. Finally, in 2000 the network of national motorways was fairly complete, and 2000 also represents the year when the surveys for the SADS started. Over time, travel times were successively approximating the pattern of Euclidean distances, as some of the major topographic obstacles have been overcome. Map 2 shows this effect. This map depicts the travel times by car in 1950, centred on the alpine town of Visp. In 2000, the maximum travel time to the farthest place in the study area was 210 minutes. Any travel time that is higher than that is shown in red colours, representing how much longer travels took in 1950 compared to 2000. It should be noted that the majority of the population actually lives in the lowlands of the Swiss Plateau, which has a much denser transport network, with less impact of surface topography.

Swiss German syntax data is only available for the year 2000 (from the SADS), apart from some limited questions in former surveys (Sprachatlas der deutschen Schweiz (SDS) - Hotzenköcherle, Schläpfer, Trüb, & Zinsli, 1962-1997), and thus absent for the other dates for which travel time data is available. For the purposes of this study, we assume that syntax changes at a slower rate over time than other linguistic levels (Longobardi & Guardiano, 2009), which means we expect historical language contact possibilities to be represented in today's syntactic landscape as well.

3. METHODOLOGY

As mentioned in the Introduction, our methodology includes the development of a measure of linguistic distance (3.1) suitable to deal with the SADS data, visualisation (3.2) and the correlation of this measure (3.3) with different geographic distances, including Euclidean distance as well as travel times. Furthermore, we conduct various analyses to reveal local variation (3.4).



Map 2 Travel times by car in 1950 from Visp. Travel times that are longer than the ones that exist in 2000 are represented with a red colour scale, representing the development of 50 years in travel times

3.1. Calculating Syntactic Distance

Linguistic (dis)similarity in syntax data has often been computed using the Hamming distance (Spruit, 2006) or Goebl's (1982) Weighted Identity Value (Spruit et al., 2009). These measures define differences between two variants, assuming one variant per survey site. Since in the SADS multiple variants may occur per survey site, we cannot assume the linguistic distance between two survey sites to be equal to the difference between the two variants.

In our study the occurrence of each variant at a given survey site is converted to ratios of the number of respondents using the particular variant divided by the total number of respondents answering at the given site. To calculate the syntactic differences for a given variable (i.e., survey question) for a particular pair of survey sites, the proportions of answer variants at the two survey sites are subtracted from each other, resulting in a difference for the given survey site pair and given variant. The overall, aggregate syntactic distance is then calculated by adding up the differences for the variables of choice.

Figure 1 shows this procedure for two simplified survey questions (SADS questions QI.01., QI.03.) and two survey sites (Klosters, Flühli). For a single variable the maximum distance between two survey sites is 2, which may be reached if the two sites do not overlap in any variants at all (e.g. survey site A uses exclusively one variant while survey site B uses exclusively another). In our case, given 60 variables, syntactic distances for a survey site pair will range on a scale of 0 to 120. An equivalent method was used before by Speelman et al. (2003) for quantifying difference between language profiles and Pickl et al. (2014) to calculate linguistic distance between survey sites. In both works the resulting sum is divided by 2 to account for bidirectionality, but in our case this is not needed if only the linguistic distance is calculated.

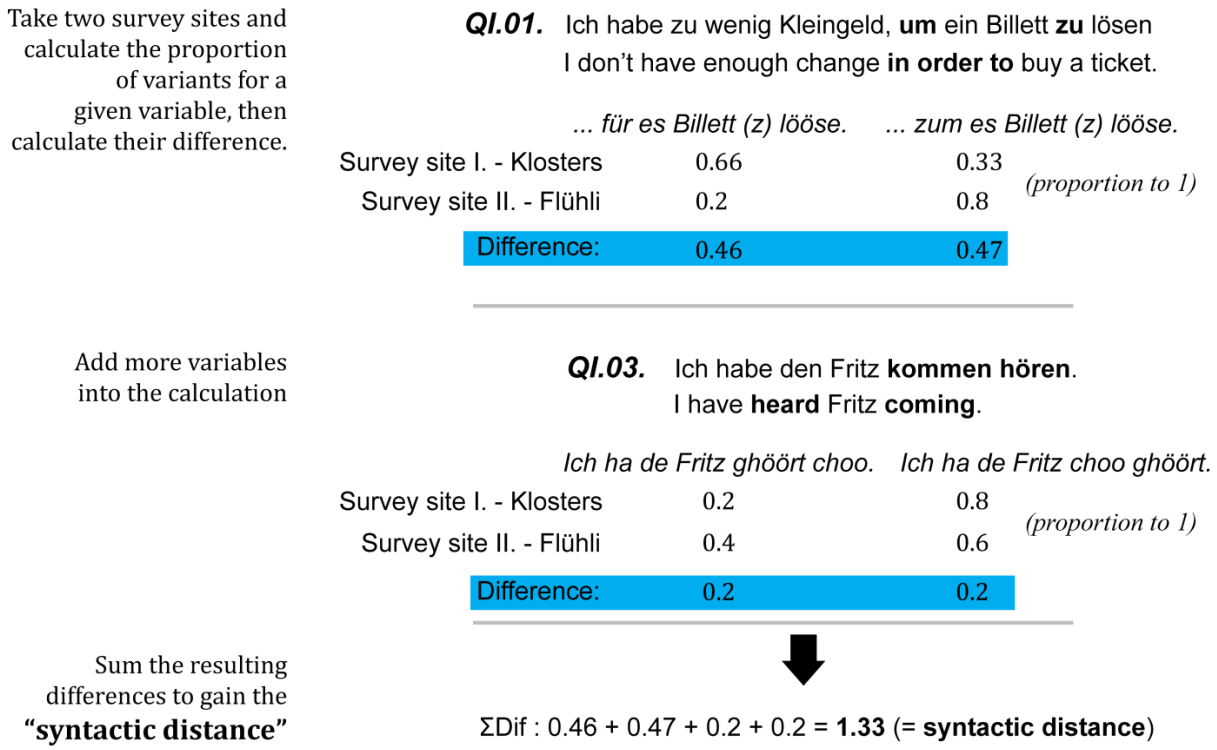


Fig. 1 Procedure to calculate the syntactic distance. Normally there are more than two answer variants for each question.

3.2. Visualisation of Syntactic Distances

The results of this syntactic distance calculation can already reveal a lot about the relationship of geographic location and language variation, simply by visualisation, comparable to maps of Goebl's identity values (e.g. Goebl, 2010). To this end, we create maps of syntactic distances centred on particular survey sites both for the entire study area as well as local spatial subsets, and a global map of average syntactic distance per site.

3.3. Correlation of Syntactic and Geographic Distances

Once the pairwise syntactic distances have been computed for all survey site pairs in the dataset, it is then possible to compute the correlation between the linguistic and the geographic distances over all survey site pairs. We use three correlation measures: two measures of linear correlation — the Pearson product-moment correlation coefficient and the distance-oriented Mantel test (Mantel, 1967), similarly to Scherrer (2012), Haynie (2012) and Grieve (2014)— and logarithmic correlation. Furthermore, regression models of different types (linear, logarithmic) are fitted to the distributions of syntactic distance against geographic distance, separately for each type of geographic distance measure used. In most former dialectometric investigations not directed at syntax (e.g. Heeringa & Nerbonne, 2001; Nerbonne, 2009; Nerbonne, 2010; Pickl et al., 2014) a logarithmic model better described the relationship between the linguistic and geographic distances, respectively. On the other hand, Spruit (2006), using syntactic data of Dutch dialects, found better agreement with linear correlation and Stanford (2012:273) found that the "patterns of dialectometry" in general do not necessarily apply in smaller areas.

3.4. Local Analyses

In order to further study the morphosyntactic variation at the local level, we carry out three types of local analyses. They enable the study of potential barrier and contact effects at the local level. Also, they allow the comparison of patterns of syntactic variation between different survey sites. The first two of these local analyses are identical to their global counterpart, but restricted to

local subsets of the study area: First, we create maps of syntactic distances for local subsets, centred on a particular survey site and second, we carry out a correlation analysis for the subsets using the same linear and logarithmic methods as at the global level. In order to explore the effect of topography, we use spatial subsets in a mountainous area and in an area with gentle topography, respectively. The third analysis is local in the sense that it allows highlighting local deviations from the global regression model using geographic distance as a predictor of linguistic difference. To this end, we compute the residuals of syntactic distance and geographic distances (Euclidean, travel time) (Section 3.5), again centred on a particular survey site.

3.5. Residuals of Syntactic and Geographic Distances

If the values of the syntactic distances as well as the geographic distances are both normalised to the interval [0..1], the differences (i.e. residuals) between the syntactic and the geographic distance values can be calculated for each survey site, in relation to a reference site. These residuals can then be visualised either in scatterplots or in area-class maps for each particular reference site. The residuals are indicative of the degree of agreement between the syntactic and geographic distances.

3.6. Implementation

The statistics software R was used for the computation of syntactic distances, correlation, regression analysis and statistical testing, and for producing diagrams. Packages *plyr* and *ggplot2* were used for preparing and plotting the data, while *ade4* was used for the Mantel test. The GIS software ArcGIS was used for producing the maps.

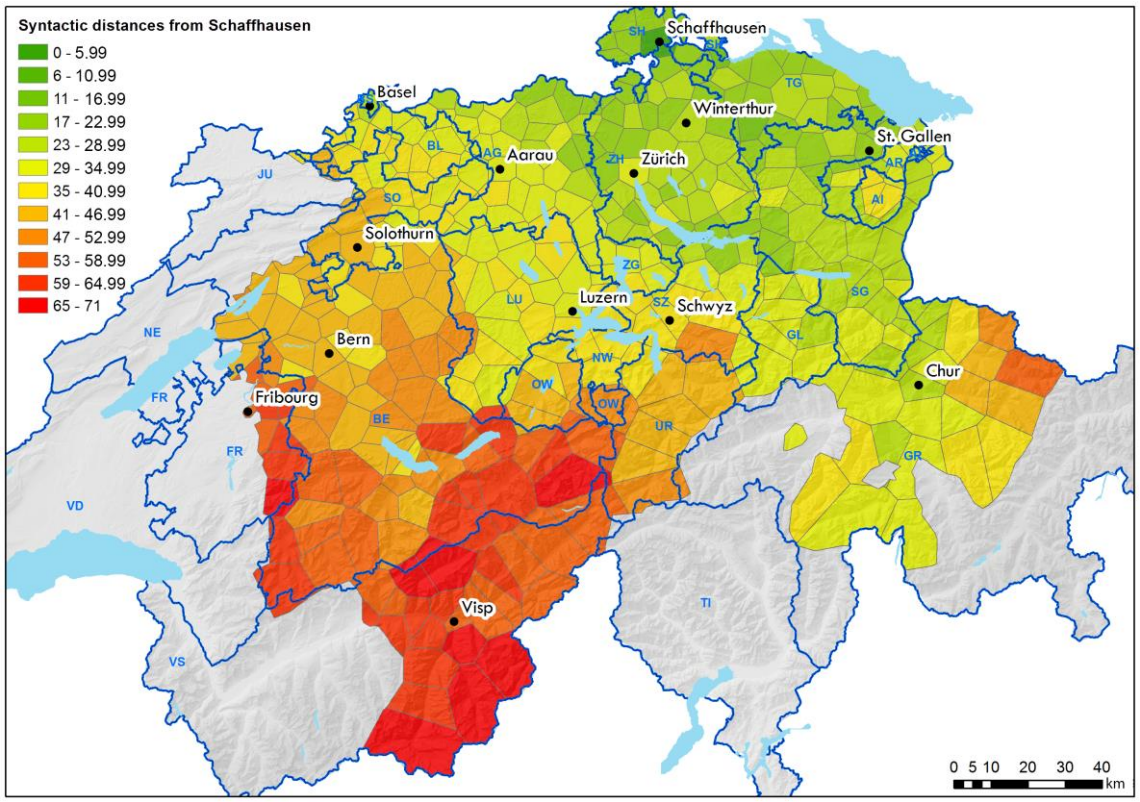
4. RESULTS

The maps presenting the results of this study feature two areas. First, the whole investigation area of the SADS surveys (383 survey sites), which roughly covers the German-speaking area of Switzerland (referred to as the *global* area). Second, one of our specific *local* subsets, consisting of 46 survey sites in the Swiss cantons Berne and Valais (*German: Wallis*), respectively. It is thus referred to as *BEOV* – short for "Bernese Oberland and Valais". This local region is dissected by a major ridge of the Swiss Alps, forming the border between the two cantons and an important topographic barrier. The Bernese part of this local subset represents the alpine part of the canton of Berne, also known as the *Bernese Oberland*. This area is characterised by a network of many valleys that are partially deep and secluded. The part in canton Valais, located to the south, is dominated by a large valley (the upper part of the *Rhône Valley*), with some side valleys, most importantly the *Lötschental* valley.

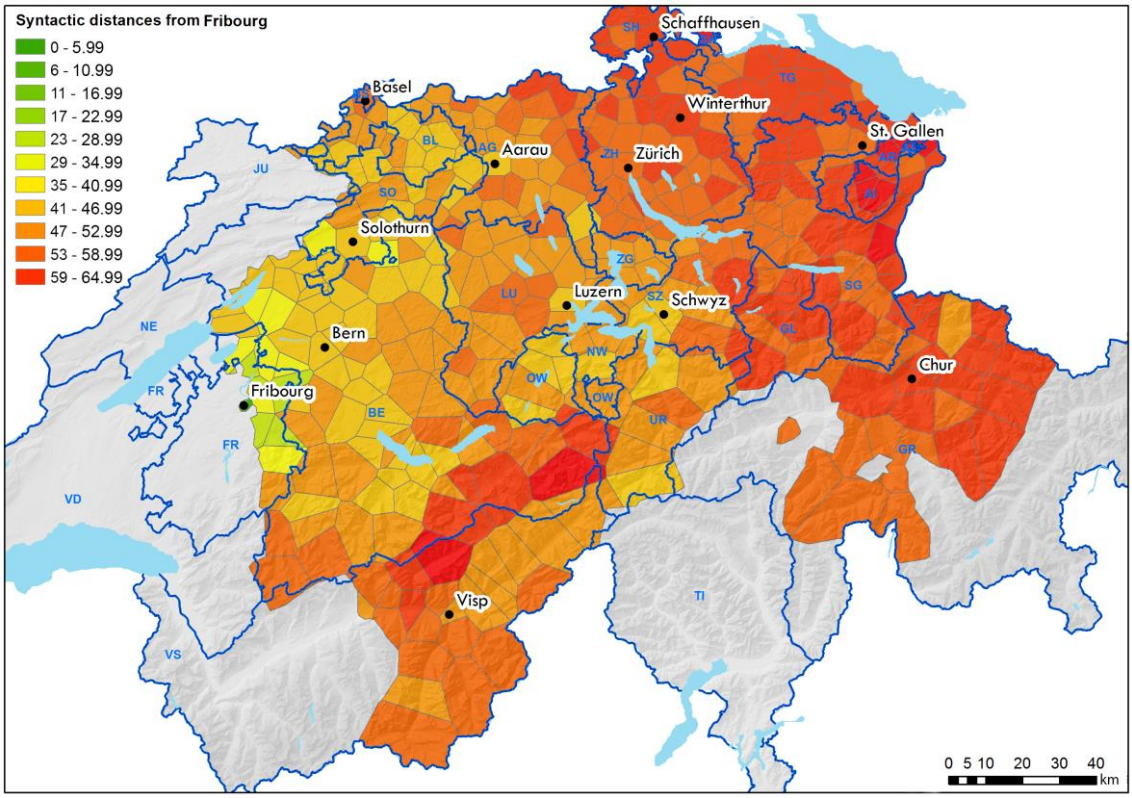
The results for a third data set — a local subset of 46 survey sites located on the Swiss Plateau (*"Mittelland"*) between the cities of Aarau, Solothurn and Berne — are presented only for the local correlation analysis (Section 4.4), and not shown in maps. This data set, with the short name *ML46*, serves as an example with gentle and homogenous topography, which also allows for better transport connections, therefore elevated possible direct contact throughout the area.

4.1. Maps of Syntactic Distance

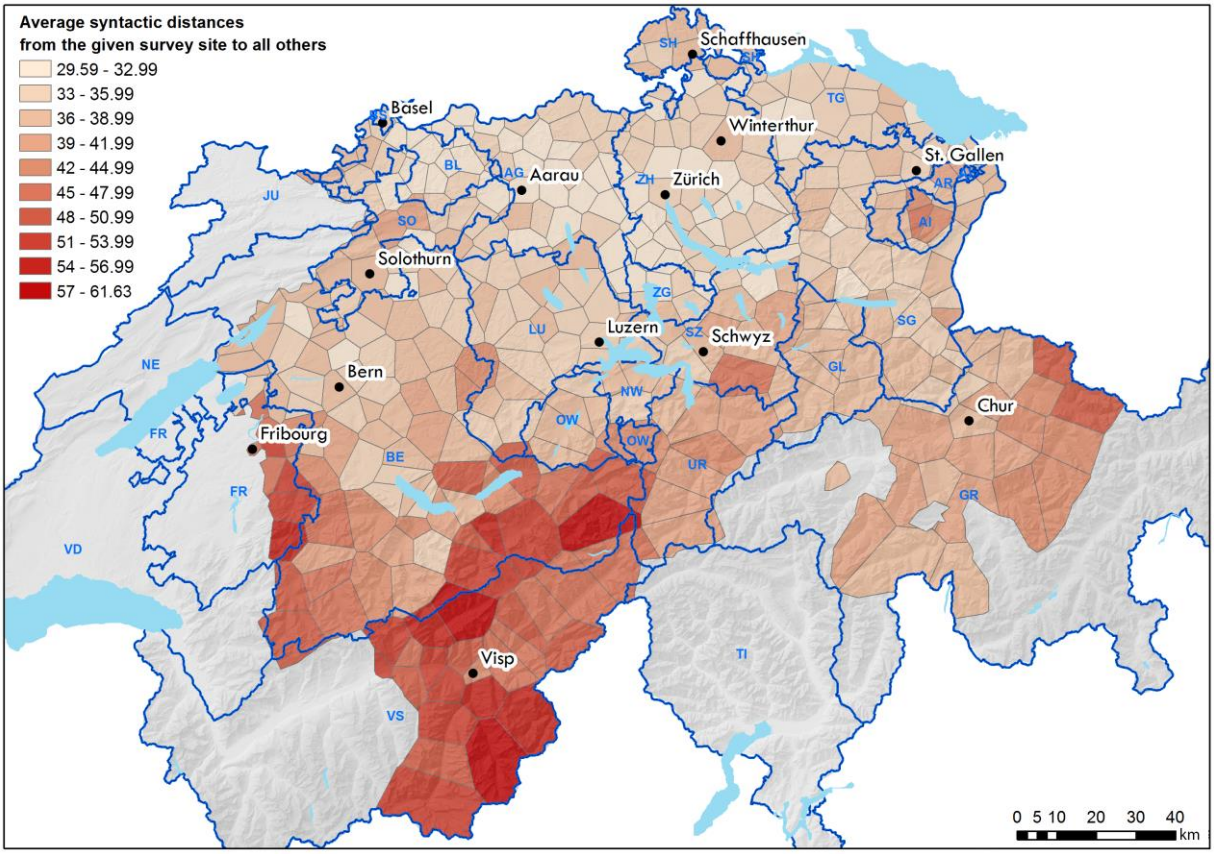
Maps 3 and 4 show the survey sites of the entire study area as Voronoi polygons (similarly to e.g. Goebl, 1983, 2010) coloured according to their syntactic distance from a particular reference place, with the borders of the Swiss cantons overlaid. We chose two places in Switzerland to present different spatial patterns of syntactic distance variation over the topographically diverse investigation area. Schaffhausen, a city on the Swiss Plateau, close to the German border, serves as reference site in Map 3, while the city of Freiburg (*French: Fribourg*), which is located at the far west end of the German-speaking area and actually has a French-speaking majority, is the anchor in Map 4.



Map 3 Syntactic distances of the survey sites compared to Schaffhausen.



Map 4 Syntactic distances of the survey sites compared to Freiburg (French: Fribourg).



Map 5 Average syntactic distances to all other survey sites.

Map 5 depicts for each survey site the average syntactic distance to all other survey sites, thus representing how different the given survey site is from all others, the darkest coloured polygons indicating the most dissimilar dialects.

4.2. Scatterplots and Correlation Analysis

Figure 2 depicts the syntactic distances plotted against the Euclidean distance for all the survey site pairs. The survey sites included in the BEOV subset are highlighted in green. The linear and logarithmic regression lines with regards to all survey site pairs on the one hand and to the BEOV subset on the other hand are overlaid. Figure 3 shows three scatterplots, where the syntactic distance is plotted against the travel times for the three dates, 2000, 1950 and 1850, respectively, along with the linear and the logarithmic regression lines.

Table 2 presents the results of the correlation analysis of the syntactic distance with the geographic distances for the entire study area. The table shows the coefficients (r) obtained from the Pearson product-moment correlation and the Mantel-test methods, respectively, along with the coefficients from the logarithmic correlation. The resulting R^2 from the regression analyses is also shown, indicating the extent to which the geographic distances account for the variance in the syntactic distance.

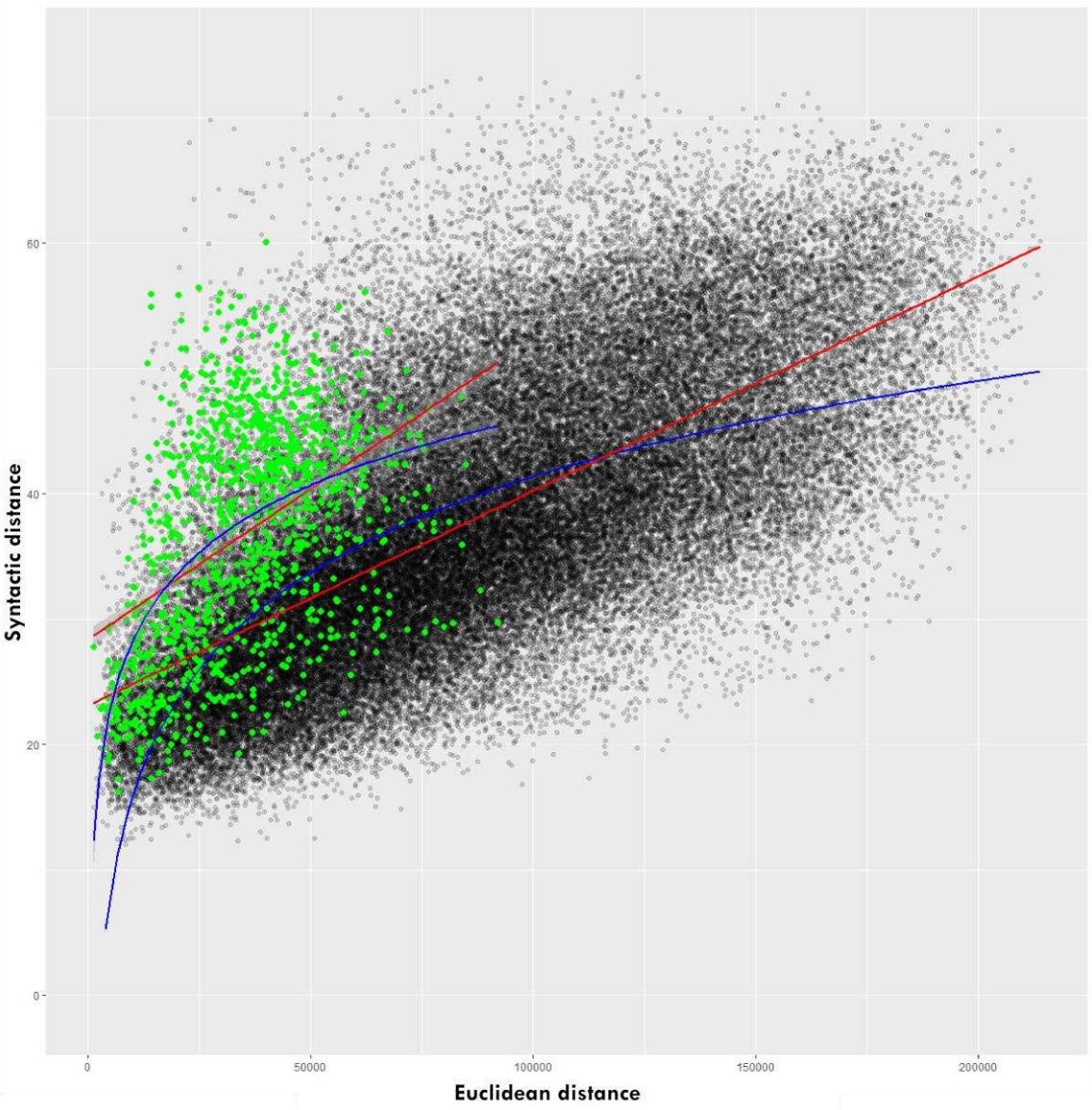


Fig. 2 Syntactic distance plotted against the Euclidean distance [m]. The survey site pairs included in the BEOV subset are highlighted in green. Linear regression line is shown in red, logarithmic regression line in blue.

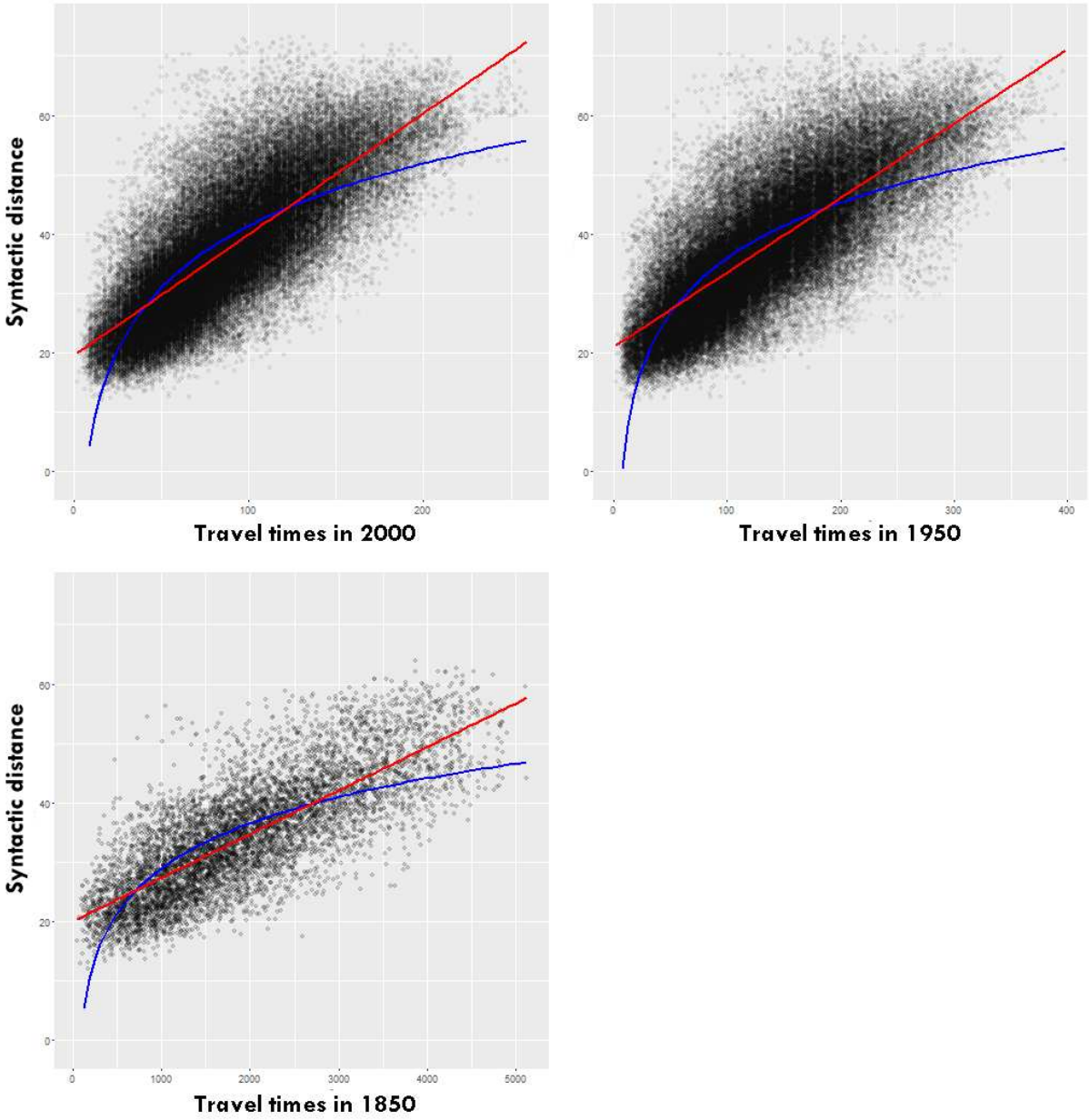


Fig. 3 Syntactic distance plotted against the travel times in minutes. Top left: against travel times in 2000; top right: against travel times in 1950; lower left: against travel times in 1850. Linear regression line shown in red, logarithmic regression line in blue.

Table 2. Correlation coefficients of the syntactic distance with the different geographic distances, as well as the explained variance R^2, for both linear and logarithmic regression analyses. All 383 survey sites. *Travel time data for 1850 was available for only 120 survey sites in the investigation area.

	Pearson's correlation		Mantel-test		Logarithmic correlation	
	r	R^2	*r*	R^2	*r*	R^2
Euclidean distance	0.676	0.458	0.65	0.422	0.65	0.424
Travel times in 2000	0.775	0.599	0.76	0.577	0.744	0.553
Travel times in 1950	0.778	0.605	0.768	0.590	0.743	0.552
Travel times in 1850*	0.783	0.612	0.763	0.582	0.737	0.544

4.3. Maps of Syntactic Distance for the BEOV Subset

Maps 6 and 7 show syntactic distances in the BEOV subset. For comparability the same colour scheme is used as for the corresponding maps of the whole study area (Maps 3 and 4). Maps 6 and 7 are centred on Blatten (Valais) and Grindelwald (Berne), respectively. Roads suitable for cars are also featured on these maps to give an impression of the main transport connections and mountain passes. In the past (i.e. 1850), similar routes were used by horse carriages and stagecoaches, further mountain passes could have only been traversed by foot or mules.

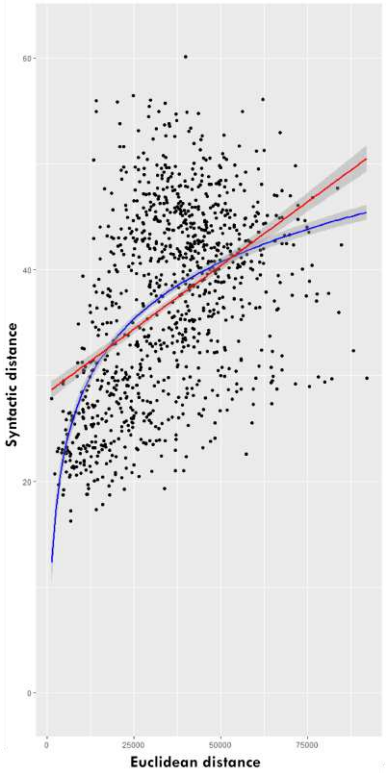
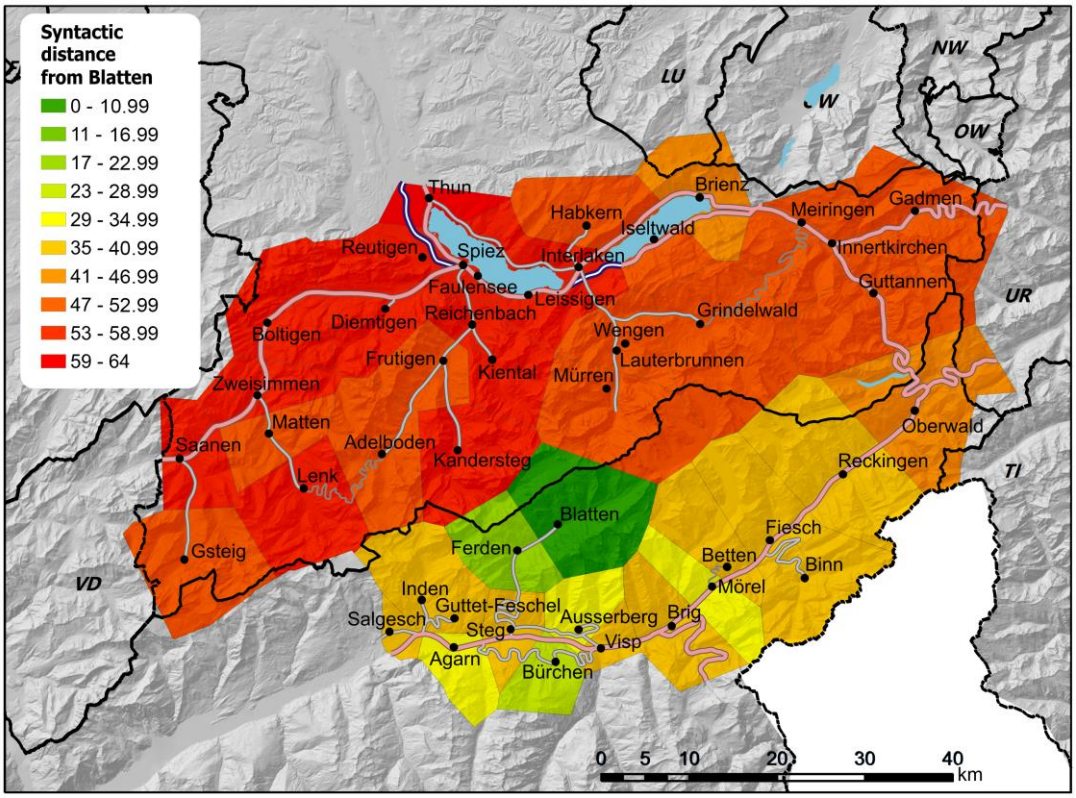
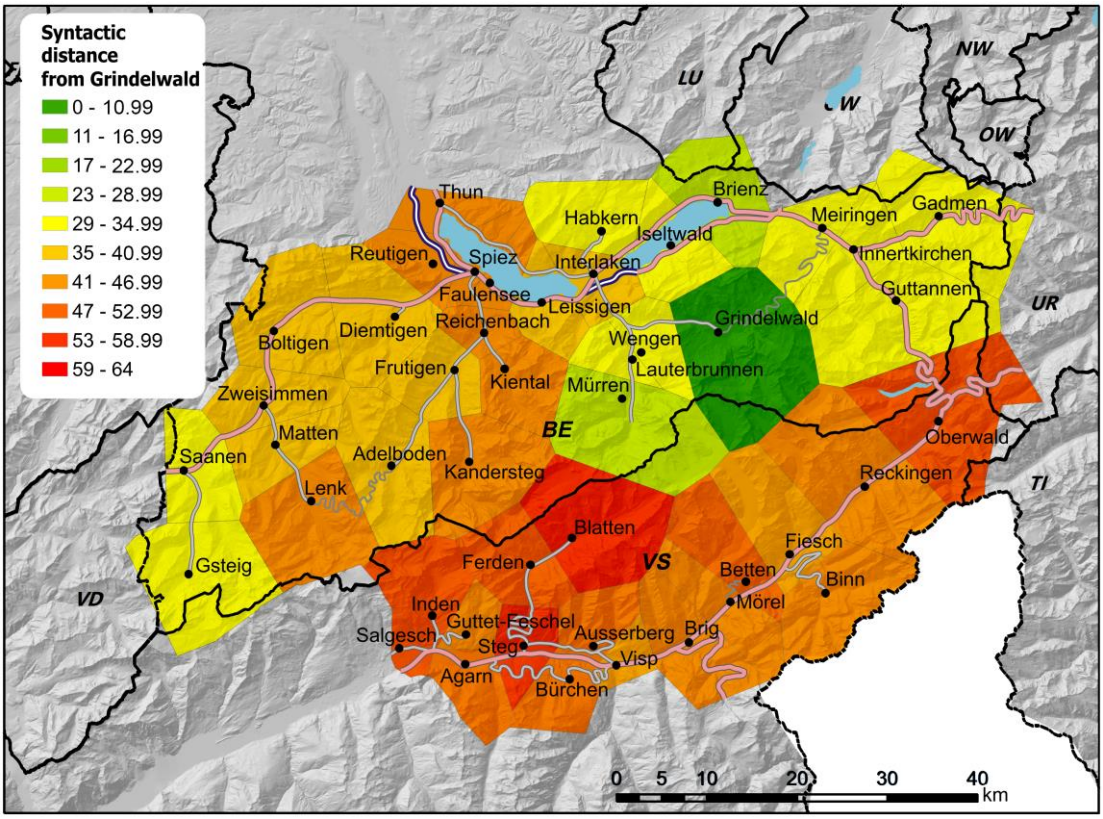


Fig. 4 Syntactic distance plotted against the Euclidean distance [m] in the BEOV subset. Linear regression line shown in red, logarithmic regression line in blue.



Map 6 Subset map centred on Blatten. Note that the Voronoi polygons used for area-class display do not respect the borders between cantons.



Map 7 Subset map centred on Grindelwald.

4.4. Scatterplots and Correlation Analysis of the Local Subsets

Figure 4 plots the syntactic distances between survey sites against the Euclidean distance while Table 3 presents for the BEOV subset the results of the Pearson correlation analysis of the syntactic

distance with the geographic distances, and with their logarithms. In the same manner, Table 4 shows the results of the correlation analysis for the ML46 subset. In order to assess whether correlation coefficients differ significantly from each other, they were tested using Fisher's *z*-transformation (Lowry, 2000; Warner, 2013), with results shown in Table 5 for the global level and the BEOV subset.

Table 3. Correlation coefficients of the syntactic distance with the different geographic distances, as well as the explained variance R^2, for the linear and logarithmic regression analyses. BEOV regional subset.
**** Travel time data for 1850 was available only for 11 survey sites in the BEOV subset.**

	Pearson's correlation		Logarithmic correlation	
	r	***R^2***	***r***	***R^2***
Euclidean distance	0.445	0.198	0.519	0.27
Travel times in 2000	0.674	0.455	0.694	0.482
Travel times in 1950	0.727	0.53	0.749	0.562
Travel times in 1850**	0.815	0.665	0.811	0.657

Table 4. Correlation coefficients of the syntactic distance with the different geographic distances, as well as the explained variance R^2, for the linear and logarithmic regression analyses. ML46 regional subset.
***** Travel time data for 1850 was available only for 19 survey sites in the ML46 subset.**

	Pearson's correlation		Logarithmic correlation	
	r	***R^2***	***r***	***R^2***
Euclidean distance	0.543	0.295	0.535	0.287
Travel times in 2000	0.577	0.333	0.557	0.31
Travel times in 1950	0.558	0.312	0.547	0.3
Travel times in 1850***	0.607	0.369	0.554	0.307

Table 5. Results of the Fisher *z*-transformation (*p*-values) to test whether the difference between correlation coefficients is significant. Significant differences ($p < 0.05$) between the two values highlighted in green.

Similarity of the Pearson product-moment correlation of the given geographic distances to the syntactic distance ; *P*-one-tailed values	Global Euclidean distance (0.676)	Global Travel times 2000 (0.775)	Global Travel times 1950 (0.778)	Global Travel times 1850 (0.783)
Global Euclidean distance (0.676)		0.002	0.0014	0.015
Global Travel times 2000 (0.775)			0.4562	0.4247
Global Travel times 1950 (0.778)				0.4522
Global Travel times 1850 (0.783)				

BEOV subset **Similarity of the Pearson product-moment correlation of the given geographic distances to the syntactic distance ; *P*-one-tailed values**	BEOV Euclidean distance (0.4446)	BEOV Travel times 2000 (0.6738)	BEOV Travel times 1950 (0.7268)
BEOV Euclidean distance (0.4446)		0.0571	0.0197
BEOV Travel times 2000 (0.6738)			0.3121
BEOV Travel times 1950 (0.7268)			

4.5. Residuals of Syntactic and Geographic Distances

Figure 5 plots the residuals of syntactic distance and Euclidean distance (shown on the y-axis) against the Euclidean distance of all survey sites relative to the alpine village of Obersaxen. Obersaxen was chosen as it is located in the periphery of the study area and scored a moderate average syntactic distance to all other survey sites (see Map 5 and Map 8 for Obersaxen's location). Maps 8 and 9, respectively, then map the residuals to geographic space. Map 8 does so for the residuals in Figure 5, relative to Obersaxen. Map 9 shows the residuals of syntactic distance and travel times in 1950, relative to Freiburg. The patterns of residuals may differ considerably depending on the reference site and the type of geographic distance used; these maps are thus to be understood as examples.

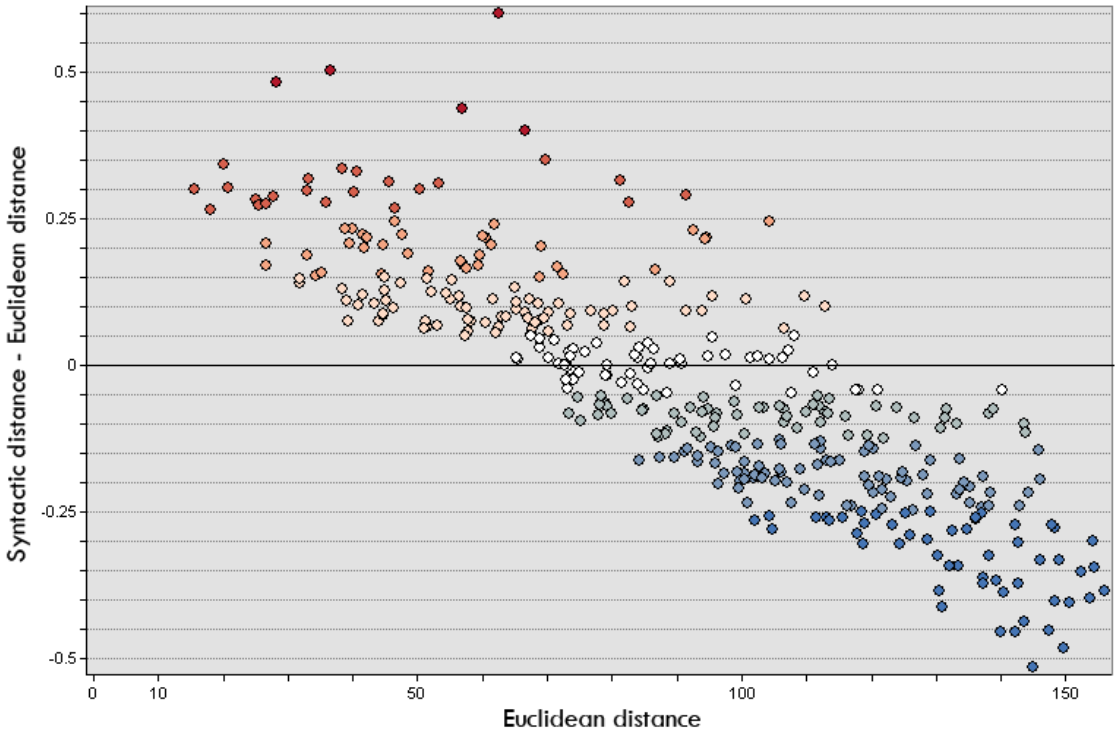
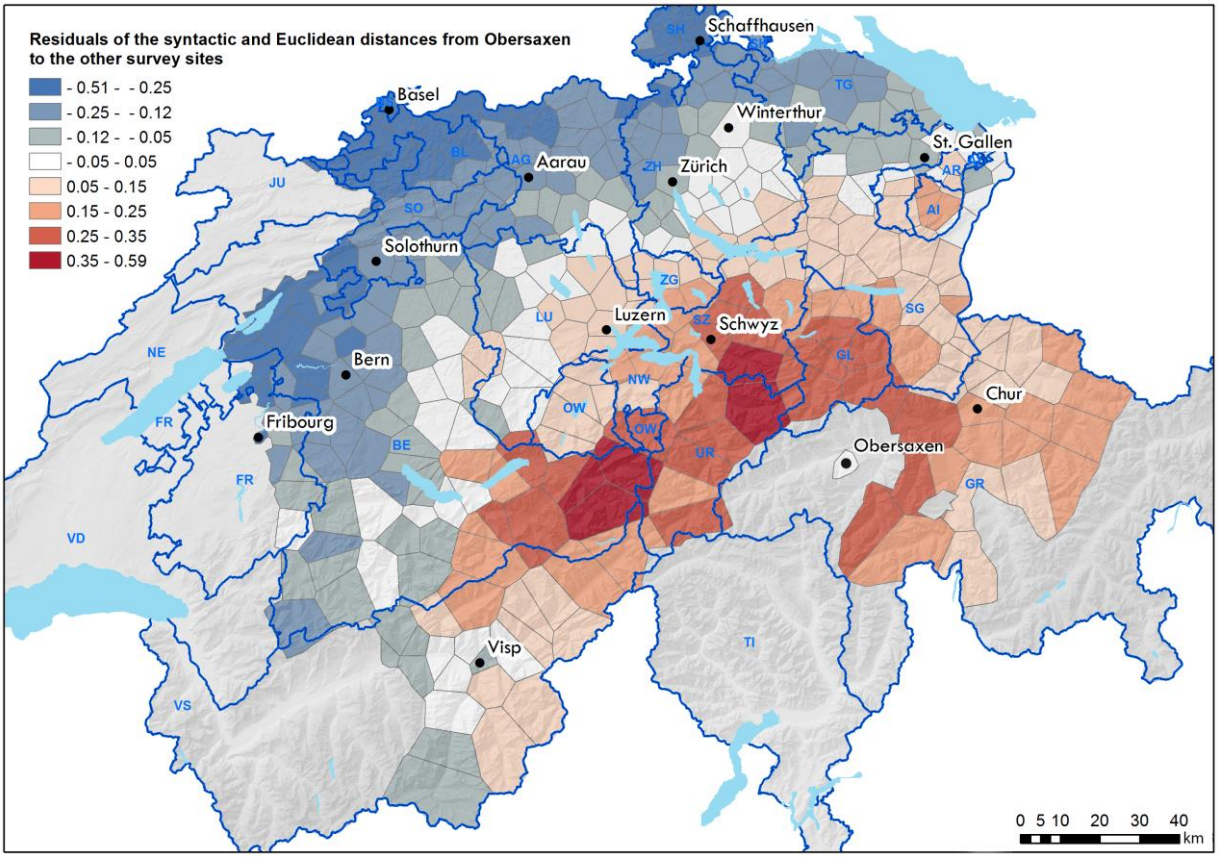


Fig. 5 Scatterplot of the residuals of syntactic distance and Euclidean distance plotted against the Euclidean distance [km] from Obersaxen to all other survey sites.



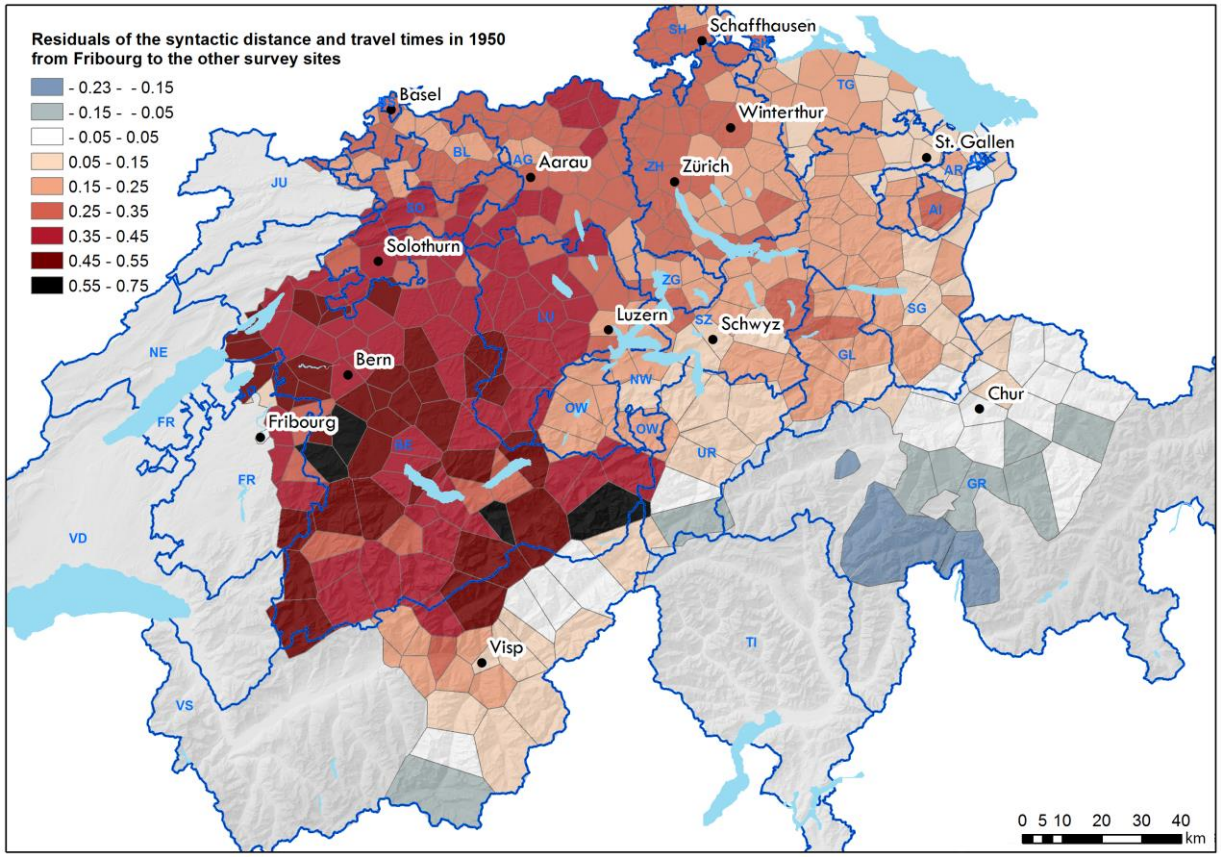
Map 8 Residual map of Obersaxen showing the residual values between the normalized syntactic distance and the Euclidean distance. If the Euclidean distance is greater than the syntactic distance, it will yield a negative residual; conversely, if the Euclidean distance is lower than the syntactic distance, a positive residual is received.

5. DISCUSSION

5.1. Syntactic Distance Measure

As explained in Section 3.1 we used an aggregative distance measure to express the linguistic (dis)similarity between survey sites, accommodating the fact that in the SADS survey multiple answers are provided per survey site. With this approach, our results are comparable to those by other authors who used similar measures in similar studies (Speelman et al., 2003; Pickl et al., 2014). However, in our measure we did not assign weights to any of the phenomena, or to single answer variants for the survey questions, although it can be assumed that between certain types of answers, some of the differences are more pronounced and that some syntactic phenomena are more salient than others, at least from a perceptional point of view. We used a linear summation, not taking into consideration the potential mutual correlations between the answer matrices of the survey questions, essentially assuming independence between the variables (survey questions).

Establishing weights for each of the phenomena and each of their answer variants, however, would be a very tedious work and prone to subjective judgement. In practice, since we are forming our aggregate measure from a large number of variables, the resulting distance values are deemed to be realistic, as the weights and uncertainties of the various variables cancel out by aggregation, as posited in Nerbonne's (2009) work.



Map 9 Residual map of Freiburg (French: Fribourg) shows the residual values between the normalized syntactic distance and the normalized travel times in 1950. For an explanation of the meaning of residuals, see Map 8.

5.2. Global Maps of Syntactic Distance

The maps of syntactic distance can be used to visualise patterns of variation of syntactic distance across the study area. Depending on the choice of the survey site on which the map is centred, the syntactic distances will show remarkably different patterns. Map 3, which is centred on Schaffhausen, exhibits a largely concentrical progression of syntactic distance values with increasing geographic distance, suggesting support of the FDP. However, in that same map, we can also see breaks in this progression, particularly at canton borders, which often also form old historical and cultural (i.e. denominational) borders. The second effect that we may perceive is the influence of topography. For instance, in Map 3 the progression of syntactic distance values is flatter in the lowlands of the Swiss Plateau (cf. Map 1), and steeper towards the more mountainous areas, except for the upper Rhine valley, which provides good accessibility towards the city of Chur. This topographic effect, we would think, should also show in the correlation analysis, where we would expect higher correlation values for the travel time distances, as opposed to the Euclidean distance. Map 4, centred on Freiburg, shows a somewhat patchier pattern than the previous map. The two effects — canton borders and topographic effect — appear less pronounced, except for a relatively stark difference along the border between the cantons Fribourg (*German: Freiburg*) and Berne, which also forms a strong denominational border (catholic vs. protestant). Clearly, the progression of syntactic distance with increasing geographic distance is steeper in Map 4 than in Map 3, suggesting that Freiburg is syntactically more different from other survey sites than is the case for Schaffhausen.

Map 5 is depicting the average syntactic distances from each survey site to all other sites. Owing to the fact that average values are shown, we now see a different picture, but some of the effects visible in the preceding maps are still noticeable. The effect of topography becomes noticeable in two ways. First, differences in average syntactic distances on the Swiss Plateau are only rather subtle, hinting at better possibilities of communication. And second, the values are generally higher in the Alps, reaching peak values in the Bernese Oberland and the Valais, areas characterised by high

mountains and topographic barriers, which foster isolation. The high values in canton Fribourg further support the observation that Freiburg is syntactically distinct from the other Swiss dialects. The rather clear discontinuity between cantons Fribourg and Berne might be less an effect of topography (which is little pronounced in this border area) than it might be caused by isolating denominational differences between a catholic canton (Fribourg) and a protestant (Berne) canton.

5.3. Global Scatterplots and Correlation Analysis

Scatterplots of geographic distance against syntactic distance graphically show the strength and the direction of correlation between the two variables, and they also allow fitting regression lines to the data points. As mentioned in Section 3.3, we use linear and logarithmic regression. Figure 2 exhibits a positive correlation between syntactic distance and Euclidean distance for the entire SADS data set. As posited by the FDP, syntactic distance grows with increasing Euclidean distance. This is further underlined by the linear and the logarithmic lines of best fit, however it is hard to tell solely by looking at the figure whether the linear regression line or the logarithmic line fits better (which contradicts findings of earlier studies, e.g. Heeringa & Nerbonne, 2001; Nerbonne, 2009; Nerbonne, 2010; Pickl et al., 2014, where the sublinear patterns were unequivocal). Although the correlation is obviously not perfect, as the data points deviate considerably from the regression lines, nevertheless a rather elevated strength of correlation is visible. The green points in Figure 2, highlighting the survey site pairs included in the BEOV subset, allow for making two qualitative observations. First, the syntactic distances in this subset tend to be higher than for other sites at equal Euclidean distance. And second, the syntactic distances appear to show a greater degree of variation, thus suggesting that the strength of correlation might be smaller in the BEOV subset than in the entire study area (cf. Section 5.5).

Figure 3 plots syntactic distance against the travel times in the years 2000, 1950 and 1850, respectively. Note that the length of the horizontal axis has been adjusted so that it fits the range of the travel times for the three reference years. Hence, each of the three subfigures of Figure 3 shows a different range of travel times in minutes, but on the same graphical length of the scatterplot, thus alleviating comparability. Visually, one gets the impression that the strength of correlation and thus the fit to the regression line might be higher than in Figure 2. The direction of correlation remains the same.

The numerical correlation analysis for the entire SADS data set, summarised in Table 2, brings about further insights. The correlation coefficients of the syntactic distance with the geographic distances at the global level always show a strong positive relationship of at least 0.65 for both linear methods of correlation (i.e. Pearson product-moment correlation and the Mantel test). The R^2 are between a low of 42.19% of explained variance of syntactic distance for the Mantel test and the Euclidean distance, and a high of 61.24% for the Pearson correlation and the travel times of 1850. All linear correlation coefficients are significant ($p < 0.05$) or highly significant ($p < 0.01$), independently of the correlation measure used. The best correlation in the linear case is obtained for the travel times of 1850, with the other travel times showing similar values. However, there appears to be a marked drop in the correlation strength when the Euclidean distance is used. In Section 5.5, we will analyse this difference in more detail, but we already note that these results hint at the topographic effect, which translates directly into travel times.

There is always a positive correlation with the logarithms of the geographic distances too, and they are in fact higher than the patterns commonly reported in dialectometric studies on pronunciation and phonology (Nerbonne's comparative study (2010) accounts for correlation coefficients between 0.469 and 0.622 using Euclidean distance). In our case, a linear model describes the relation between linguistic and geographic distances slightly better than a logarithmic model (the maximum difference between R^2 values is 6.8%, for the 1850 travel times), but the difference is in no case statistically significant. This means that for our syntactic distance both the linear and the logarithmic model are equally good predictors. This finding aligns with the results of Spruit (2008),

who found that the relationship between syntactic and geographic distances could be slightly more accurately described with a linear function than with a logarithmic transformation. The Mantel test results are also very similar to the Pearson correlation coefficients; their difference is also in no case statistically significant. For both linear and logarithmic cases, the Euclidean distance yields markedly lower correlation values than the travel time measures. On the other hand, the monotonic progression visible in the linear correlation measures for the travel times in 2000, 1950 and 1850, is no longer present for the logarithmic measures.

5.4. Maps of Syntactic Distance for the BEOV Subset

To explore the correlations of the linguistic and geographic distances at the local level, the syntactic distances were also mapped for multiple subsets, with Maps 6 and 7 serving as examples on the BEOV subset. These maps also present the main roads, which provide the major routes of modern contact in this mountainous area. The major topographic feature in this area is the high mountain chain that forms the border between the cantons Berne and Valais. The Grimsel Pass between Guttannen and Oberwald, in use since the Middle Ages and open for cars, as well as the Lötschberg railway tunnel between Kandersteg and Ferden, opened in 1913 (with a new base tunnel in operation since 2007), provide connections across this topographic barrier.

Maps 6 and 7 use the same construction principle and colour scheme as the syntactic distance maps depicting the entire study area (Maps 3 and 4), and are centred on Blatten (Map 6) and Grindelwald (Map 7), respectively. Although the two maps show different patterns, we can clearly see the effect of the main mountain chain, which acts as a linguistic divider in both maps. The village of Blatten, the reference in Map 6, is located at the back end of the Lötschental valley, a secluded side valley whose entrance from the main valley of the Valais is formed by a ravine with a vertical drop of some 700 metres, and which is separated from the canton of Berne by mountains exceeding 3000 and even 4000 metres. This particular location suggests how topography exerts an influence on linguistic differentiation. Ferden, which is shown to be syntactically close to Blatten, is located in the same valley, while there is already a marked difference visible to the other survey sites in the Valais, and a clear-cut difference to the sites across the mountain chain in the canton of Berne. In general, the farther we follow the transportation routes, the higher the syntactic distances that we observe, in agreement with the FDP.

Map 7 is centred on Grindelwald, a rather well accessible valley in the Bernese Oberland attracting many tourists. At first sight, the patterns we see are similar to those shown in Map 6, only mirrored along the main mountain chain. However, at closer inspection, we see a different form of spatial variation, influenced by (so we believe) a different topography. Rather than a main valley with one major side valley, as in the case of the Valais, in the Bernese Oberland we find several separate valleys, as is clearly reflected in the road network. These valleys have quite different syntactic distances from Grindelwald, in some cases even reversing the colour scale with increasing geographic distance, and thus contradicting the FDP.

We also explored the potential ‘bridging’ effect of the Grimsel Pass and the Lötschberg tunnel. This effect would suggest that the syntactic difference between survey sites connected by a pass or tunnel should be smaller than between other sites. However, when we generated centred syntactic difference maps for all survey sites (the complete series not shown here), we found this effect to be rather weak and unsystematic.

Concluding the visual part of our analysis of the BEOV subset, we observe a mixed picture. On the one hand, topography appears to have a strong effect on linguistic differentiation, as shown very clearly through the main mountain chain, and also in the way survey sites of the same valley tend to be more similar than their neighbours in the next valley. This separating effect of topography was also observed in perceptual linguistics studies, where laypersons drawing dialect areas would intuitively link these to topographic features (Stoeckle, 2014:369, 519). On the other hand, we also see deviations from this pattern, as seen on the example of Grindelwald in Map 7, suggesting that other factors than geographic distance and topographic effects come into play.

5.5. Scatterplots and Correlation Analysis for the Local Subsets

Comparing Table 2, representing the results of the correlation analysis at the global level, and Table 3, representing the results for the local BEOV subset, we notice that the situation is quite different for the two data sets. With the exception of the travel times in 1850, correlation strength is higher for the global data set (Table 2) than for the local BEOV subset (Table 3). In Table 3 as well, syntactic distance correlates more strongly with travel times (which incorporate topography) than with Euclidean distance and that difference has become more marked, compared to the global data set (14.1% vs. 25.7% more variance explained by travel times). Lower correlation coefficients with Euclidean distances were indeed expected (as already suggested in the discussion of Figure 2), given the fact that throughout the BEOV region, topography crucially influences possibilities of contact. We further see that the more we move back in time, the correlation is monotonically increasing up to 0.815, suggesting that the separating effect of topographic barriers is decreasing as new, better transportation infrastructure is built. Note however, that travel times in 1850 were only available for 11 of the 46 survey sites in the BEOV subset.

In contrast to the global correlations represented in Table 2 (except for the case of 1850), in the BEOV subset the logarithmic correlation coefficients are higher than the linear Pearson correlation coefficients (Table 3). Differences between the respective correlation coefficients not being statistically significant however means that also at this local level a logarithmic model is equally good for describing the relation between syntactic and geographic distances as the linear model. This fact is also supported by Figure 4 where it is unclear whether the linear or the logarithmic regression line has a better fit. As Mantel-test results on the global level were not significantly different from the Pearson correlation coefficients, we did not include these results for the local subsets.

To further explore the impact of topography, the correlation analysis was also conducted for the ML46 subset, which features a very gentle topography, with results shown in Table 4. This subset aims to model the maximum direct language contact possible in a coherent spatial subset. The correlation coefficients obtained are very similar for the linear and logarithmic Pearson correlations and all geographic distance measures. In all but one instance, they are between 0.5 and 0.6, reaching a high of 0.607 for the Pearson correlation using the 1850 travel times. The values obtained for the Euclidean distance are the lowest for both correlation methods, however with very subtle differences to the travel time correlations. It is thus not surprising that none of the differences in correlations, when tested using Fisher's *z*-transformation, came out as significant. We conclude that in this part of the study area, characterised by a dense transportation network with no considerable topographic barriers, the effect of topography — represented by the difference between Euclidean distance and travel times — does not play out as much as on the global level, and not nearly as much as in the mountainous BEOV subset.

In turn we can also assume that the elevated level of contact possibilities also lends more opportunity for other, mainly socio-demographic variables to impact the linguistic differences, leading to lower correlations with geographic distances. When including all survey sites in the analysis, the majority of distances between survey site pairs are too big to have direct linguistic contact present. Thus, we posit that the effect of socio-demographic variables is suppressed by the effect of distance, resulting in higher correlations with geographic distances on the global level, than in local subsets. To test the assumption that geographic distance has a greater effect if no direct contact is possible, we analysed another subset of survey sites (termed *Edge46*, also *N*=46) where presumably little to no direct language contact is present. We systematically sampled the survey sites at the edge of the investigation area to simulate the distribution with the farthest distances possible. We assume that in this subset we can model the effect of geographic distances, undisturbed by direct contact (thus as clear of socio-demographic factors as possible). The resulting correlation coefficients (Table 6) are remarkably similar to those obtained at the global level (Table 2); logarithmic correlation coefficients are lower than the linear Pearson correlation coefficients. This

suggests that geographic distances have similar effects at the global level as they have in the Edge46 subset, simulating elevated isolation. The results are in contrast with the findings in the ML46 subset, which features lower correlation coefficients in all cases, and where we assume more potential direct language contact.

Table 6. Correlation coefficients of the syntactic distance with the different geographic distances, as well as the explained variance R^2, for the linear and logarithmic regression analyses. Edge situation subset (Edge46). ** Travel time data for 1850 was available only for 10 survey sites in the edge situation subset.**

	Pearson's correlation		Logarithmic correlation	
	r	R^2	*r*	R^2
Euclidean distance	0.692	0.488	0.675	0.456
Travel times in 2000	0.775	0.601	0.752	0.566
Travel times in 1950	0.775	0.600	0.750	0.563
Travel times in 1850****	0.734	0.539	0.687	0.472

5.6. Evaluating the Hypotheses

In Section 1.1, we formulated three hypotheses for this work, H1 to H3. We will now discuss each of these hypotheses in turn.

H1 states that geographic distance explains the majority of the variance found in Swiss German syntax, as represented in the SADS data. However, $R^2 > 0.5$ really only holds for the travel times at the global level (Table 2), while in the case of Euclidean distance, the threshold of the coefficient of determination for both linear and logarithmic correlations is missed (R^2 = 45.78%, 42.18%, 42.35%). In the case of the more mountainous BEOV subset (Table 3) the Euclidean distances clearly did not reach the threshold (R^2 = 19.77% and 26.96%), while travel times in 2000 only slightly missed it for both the linear and the logarithmic case (R^2 = 45.46% and 48.2%). At the same time for the ML46 subset featuring gentle topography, geographic distances in no case explain the majority of variance (Table 4).

Despite the fact that some of the R^2 fell below the 50% threshold, the strength of correlation is considerable, with R^2 often reaching values greater than 60% (Tables 2 and 3). Particularly when comparing to the results of Szmrecsanyi's (2012) morphosyntactic study of English dialects, we obtained much higher coefficients of determination. This could be due to the different data source used: Szmrecsanyi used a frequency-based casual corpus dataset, while the SADS is a survey-based atlas aimed to discover the syntactic variation as deeply as possible, with a tendency of choosing phenomena that were assumed to show spatial variation patterns. It also might be due to the fact that Szmrecsanyi's data had much coarser spatial granularity (available for the former counties of Great Britain). The strength of correlation in our case is also higher than Spruit's (2006, 2008) findings for Dutch syntactic differences using the SAND Atlas data (Barbiers, Bennis, de Vogelaer, Devos & van der Ham, 2005), where R^2 stays below 50%.

H2 posits that travel time measures better reflect syntactic spatial variation than Euclidean distance. Qualitatively, this hypothesis is very clearly supported by our results, as the correlation and determination values are always higher for the travel times than for the Euclidean distance. As shown in Table 5, the differences between Euclidean distance and travel times are also in all cases *statistically significant* for the global data set, and in one of two cases for the BEOV subset, with the other case just barely missing the 95% confidence threshold. The fact that the significance is lower for the BEOV subset than for the global dataset, despite the differences of correlation values being larger, can be explained by the much lower number of observations in the BEOV subset (N=46),

compared to the entire study area (*N*=383). The effect of gentle topography inducing elevated transportation and communication possibilities in the ML46 subset is reflected in the overall lower correlation of geographic distances with the syntactic difference. Not surprisingly, travel times in this subset are not significantly better at explaining the syntactic difference, statistically speaking (thus values are not shown.)

The correlation analysis clearly supports H2 on the global level and in the more mountainous BEOV subset. These findings are in agreement with Szmrecsanyi's (2012) observation that geographic distance *per se* does not explain the vast majority of the syntactic variation. Geographic distance is only a proxy of potential language contact or isolation, which presents itself nicely in the fact that travel times — which better reflect the actual effort that needs to be spent in order to establish contact — yield higher correlation values than Euclidean distance.

H3 states that older travel times better represent syntactic spatial variation. Based on Table 3, this hypothesis seems to be clearly supported; in Table 2 however, only the Pearson correlation values are systematically increasing from 2000 to 1950 to 1850. For Table 4, as mentioned above, there is no monotony to be observed. As seen in Table 5, differences between the correlations of different travel times with the syntactic distance are not significant. This means that, statistically speaking, for variation in Swiss German syntax, as represented in the SADS data, travel times are predictors of equal power, regardless of what year is taken.

5.7. Residuals of Syntactic and Geographic Distances

We recall that in our case the residuals are not residuals of a regression analysis; they are instead obtained as the difference of the normalised syntactic distance minus the normalised geographic distance, centred on a particular reference site (Section 3.5). Positive residuals mean that the normalised geographic distance is smaller than the corresponding syntactic distance; negative residuals indicate the opposite relation. The relationship of syntactic and geographic distance can also be understood as a simple linear regression model with a single predictor variable, geographic distance. Positive residuals would then suggest that geographic distance underestimates syntactic difference between two survey sites, while negative residuals would suggest overestimation.

Figure 5 plots, for the reference site Obersaxen, the residuals of the normalised syntactic and Euclidean distance (y axis) against the Euclidean distance (x axis). If the syntactic distance from the reference site Obersaxen was in perfect linear agreement with the Euclidean distance, no residuals would show in this graph. The residuals, however, follow a decreasing (and linear) trend.

The residuals are positive at short ranges, meaning that the Euclidean distance underestimates short-range syntactic variation under the assumption that the syntactic distance will follow a growth linearly proportionate to the Euclidean distance from the reference site. Thus surrounding dialects are more different, than suggested by the Euclidean distance. The opposite is the case at long ranges, where Euclidean distance overestimates syntactic variation. This overestimation at long ranges is rational, as geographic distance increases continuously, whereas the syntactic distance may only increase to a certain level. If two dialects become too dissimilar, they will be considered two different languages, as mutual intelligibility is no longer maintained. Zero residuals, which would mean perfect correlation of the syntactic and the Euclidean distance, occur mostly in the range of 70 to 110 km.

The geographic patterns that the residuals exhibit become more apparent when maps are used to depict the residuals. Map 8 presents the same residual values for the reference site Obersaxen that Figure 5 showed in a scatterplot. We can again see the overall trend of underestimation at short ranges and overestimation at long ranges. This trend pattern is rather systematic, evolving in concentric circles outwards from the reference site. Furthermore, the numerical range of residuals is almost symmetrical (-0.51 to 0.59), with about the same number of negative and positive residuals (cf. Figure 5). As Map 5 has shown, Obersaxen is fairly moderate regarding its average syntactic differences to all other dialects, which might explain this well-behaved pattern. Map 9 is centred on

the city of Freiburg (Fribourg), which in Map 5 has shown to be among the survey sites with the highest average syntactic difference to all other sites. Furthermore, travel times of 1950 have been used to produce Map 9 instead of Euclidean distances. The pattern visible in this map differs considerably from the map of Obersaxen. Positive residuals persist almost throughout the entire study area, meaning that syntactic variation is underestimated almost everywhere regardless of geographic distance. Only parts of the canton of Valais and Grisons show negative residuals. As a consequence, the numerical range and distribution of the residuals is highly skewed towards the positive values. Also, the concentric progression seen in Map 8 is almost not visible here, and high residual values are not restricted to short geographic distances. Indeed, the geographic range at which residuals decay to zero is much larger than in Map 8. We attribute the patterns visible in Map 9 to the special position of Freiburg in the Swiss dialect syntax landscape (Bucheli Berger, 2010; Scherrer & Stoeckle, 2016:109). Thus, we can claim that the analysis of residuals, both in scatterplots and even more so in maps, is an interesting tool to reveal such patterns and differences.

6. CONCLUSION

In this study, we compared different geographic distance measures (Euclidean distance and travel times for years 2000, 1950 and 1850) as an estimate of language contact possibility to a measure of linguistic difference between survey sites. To this end, we calculated syntactic distance based on survey data of the Syntactic Atlas of German-speaking Switzerland (SADS), involving multiple respondents per survey site, and computed different forms of correlation between syntactic distance and the different geographic distance measures, both for the entire SADS data set, as well as for local subsets. Furthermore, we generated different visualisations, again at the global and the local level.

The study set out from three hypotheses H1 to H3. Regarding H1, we showed that in most cases geographic distances explain the majority of variance inherent to the syntactic distance. Concerning H2, we have found that travel times are significantly better predictors for syntactic distance than Euclidean distance. Finally, regarding H3, although older travel times seem to be better predictors for the syntactic distance, yielding higher correlation values, their superior performance did not prove statistically significant.

We further extended our analysis to the local level, enabling discovery of a more differentiated picture of the dialectal variation across space. At the local level, the effect of topographic barriers and the effect of potential direct (language) contact became more noticeable both in the visual representation of the maps of syntactic distance as well as in the correlation analysis. Building on our results, we can conclude that the aforementioned ‘Fundamental Dialectological Postulate’ (FDP) (Nerbonne & Kleiweg, 2007: 154) seems to be true, especially on a global scale, that is, when little direct contact between speakers can be assumed. However, on a local scale linguistic distance (or similarity) depends much more on the particular characteristics of an area. This finding is in line with Stanford (2012: 274) who states that “the issue of geographic size appears to be related to fundamental distance relationships in human interaction”. On a local level, geographic distance may explain linguistic differences if, for example, topography is very pronounced and therefore actually poses a communicative boundary. If it is not, other factors (such as socio-demographic, cultural or attitudinal) may become more important.

At the level of the entire SADS data set, unlike in most other dialectometric research concerning other linguistic levels (where sublinear patterns were found unequivocal), a linear model described the correlations between geographic and linguistic distance better than a logarithmic model, although this difference is not statistically significant. Regarding the local subsets that we tested, logarithms of the geographic distances proved just as good a predictor for the syntactic distances, as the linear geographic distances.

Mapping the average syntactic distance to all other survey sites provided a way to find the dialects that are most different from the others in the syntactic sense, comparable to maps of Goebl's identity values (Goebl, 2010). By computing residuals of normalised syntactic and geographic distances, we provided a way to show to what extent and by which pattern geographic distances predict the syntactic differences.

In order to account for linguistic variation more precisely, a number of other geographic and demographic factors should also be taken into account. Empirically, urban centres are important in the spread of linguistic innovations and might therefore cause dialects to converge to dialects spoken in economically, politically and culturally dominant places (Chambers & Trudgill, 1998:172). Finding measures for this gravity-like effect, possibly using Trudgill's linguistic gravity index (Szmrecsanyi, 2012) would be desirable along with testing measures known in GIScience, such as cost distances or terrain roughness.

As before the spread of individual transport public transport meant the only connection to other parts of the country for many people, using travel times of public rather than individual transport might lead to an improvement. Also, we plan to investigate travel times from before 1850 for correlation with our syntactic distance measure.

Beyond the study of mere correlations at the aggregate level, ultimately it will be most interesting to compare spatial patterns of linguistic variation, such as local breaks in the dominance of dialectal variants, suggesting isoglosses, to geographic borders and extrageographical (political, cultural, historical; e.g. Hotzenköcherle, 1986) patterns.

REFERENCES

Barbiers, Sjef, Hans J. Bennis, Gunther De Vogelaer, Magda Devos & Margreet H. van der Ham. 2005. *Syntactische Atlas van de Nederlandse Dialecten/Syntactic Atlas of the Dutch Dialects Volume I*. Amsterdam: Amsterdam University Press.

Bucheli, Claudia & Elvira Glaser. 2002. The Syntactic Atlas of Swiss German dialects: empirical and methodological problems. In Sjef Barbiers, Leonie Cornips & Susanne van der Kleij (eds.), *Syntactic Microvariation*, Vol. 2. Amsterdam: Meertens Institute Electronic Publications in Linguistics. 41–73.

Bucheli Berger, Claudia. 2010. Dativ für Akkusativ im Senslerischen (Kanton Freiburg). In Helen Christen, Sibylle Germann, Walter Haas, Nadia Montefiori & Hans Ruef (eds.), *Alemannische Dialektologie: Wege in die Zukunft. Beiträge zur 16. Arbeitstagung für alemannische Dialektologie in Freiburg/Fribourg vom 07.-10.09.2008*, 71–83. Stuttgart: Steiner.

Chambers, Jack K. & Peter Trudgill. 1998. *Dialectology*. 2nd ed. Cambridge: Cambridge University Press.

Christen, Helen. 1998. Convergence and divergence in the Swiss German dialects. *Folia Linguistica* 32(1–2). 53–68. doi:10.1515/flin.1998.32.1-2.53.

Fröhlich, Philipp, Thomas Frey, Serge Reubi & Hans Ulrich Schiedt. 2004. *Entwicklung des Transitverkehrs-Systems und deren Auswirkung auf die Raumnutzung in der Schweiz (COST 340): Verkehrsnetz-Datenbank*. (COST). Zürich. http://www.ivt.ethz.ch/vpl/publications/reports.

Goebl, Hans. 1982. *Dialektometrie*. (Österreichische Akademie der Wissenschaften, Philosophisch-historische Klasse, Denkschriften, 157. Band). Wien.

Goebl, Hans. 1983. "Stammbaum" und "Welle". *Zeitschrift für Sprachwissenschaft* 2(1). 3–44.

Goebl, Hans. 2010. Dialectometry and quantitative mapping. In Alfred Lameli, Roland Kehrein & Stefan Rabanus (eds.), *Language and Space. Vol. 2: Language Mapping*. 433–457. Berlin: De Gruyter Mouton.

Gooskens, Charlotte. 2004. Norwegian dialect distances geographically explained. In Britt-Louise Gunnarson, Lena Bergström, Gerd Eklund, Staffan Fridella, Lise H. Hansen, Angela Karstadt, Bengt Nord- berg, Eva Sundgren & Mats Thelander (eds.), *Language Variation in Europe. Papers from the Second International Conference on Language Variation in Europe ICLAVE Vol. 2. 2004.* 195–206. Uppsala.

Grieve, Jack. 2014. A comparison of statistical methods for the aggregation of regional linguistic variation. In Benedikt Szmrecsanyi & Bernhard Wälchli (eds.), *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*, 1–34. Berlin/New York: Walter de Gruyter. doi:10.1515/9783110317558.53.

Griffith, Donald Atkins. 1987. *Spatial autocorrelation - A Primer*. Regional Science and Urban Economics. Washington, DC: Association of American Geographers. doi:10.1016/0166-0462(92)90032-V.

Haynie, Hannah Jane. 2012. *Studies in the History and Geography of California Languages*. University of California, Berkeley. 284 p.

Heeringa, Wilbert & John Nerbonne. 2001. Dialect areas and dialect continua. *Language Variation and Change* 13(3). 375–400.

Hotzenköcherle, Rudolf. 1986. *Dialektstrukturen im Wandel*. Aarau

Hotzenköcherle, Rudolf, Reinhardt Schläpfer, Reuel Trüb & Peter Zinsli (eds.). 1962-1998. *Sprachatlas der deutschen Schweiz*. I.-VIII. Tübingen: Francke.

Longobardi, Giuseppe & Cristina Guardiano. 2009. Evidence for syntax as a signal of historical relatedness. *Lingua* 119(11). 1679–1706. doi:10.1016/j.lingua.2008.09.012.

Lowry, Richard. 2000. VassarStats: website for statistical computation. http://vassarstats.net/rdiff.html? (20 September, 2016.)

Mantel, Nathan. 1967. The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research* 27(2). 209–220. doi:10.1038/212665a0.

Nerbonne, John. 2009. Data-Driven Dialectology. *Language and Linguistics Compass* 3(1). 175–198. doi:10.1111/j.1749-818X.2008.00114.x.

Nerbonne, John. 2010. Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 365(1559). 3821–8. doi:10.1098/rstb.2010.0048.

Nerbonne, John & Peter Kleiweg. 2007. Toward a dialectological yardstick. *Journal of Quantitative Linguistics* 14(2). 148–167.

Pickl, Simon, Aaron Spettl, Simon Pröll, Stephan Elspaß, Werner König & Volker Schmidt. 2014. Linguistic distances in dialectometric intensity estimation. *Journal of Linguistic Geography* 2(1). 25–40. doi:10.1017/jlg.2014.3.

SADS = Glaser, Elvira (ed.) (forthcoming) *Syntaktischer Atlas der Deutschen Schweiz*. University of Zurich.

Scherrer, Yves. 2012. *Generating Swiss German sentences from Standard German: a multi-dialectal approach.* Université de Genève. 446 p.

Scherrer, Yves & Philipp Stoeckle. 2016. A quantitative approach to Swiss German – Dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica* 24. 92–125. doi:10.1515/dialect-2016-0006.

Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35(138). 335–357.

Shackleton, Robert G. Jr. 2007. Phonetic Variation in the Traditional English Dialects: A Computational Analysis. *Journal of English Linguistics* 35(1). 30–102. doi:10.1177/0075424206297857.

Speelman, Dirk, Stefan Grondelaers & Dirk Geeraerts. 2003. Profile-Based Linguistic Uniformity as a Generic Method for Comparing Language Varieties. *Computers and the Humanities* 37. 317–337. doi:10.1023/A:1025019216574.

Spruit, Marco René. 2006. Measuring Syntactic Variation in Dutch Dialects. *Literary and Linguistic Computing* 21(4–Progress in Dialectometry: Toward Explanation). 493–506.

Spruit, Marco René. 2008. *Quantitative perspectives on syntactic variation.* Utrecht: LOT. 157 p.

Spruit, Marco René, Wilbert Heeringa & John Nerbonne. 2009. Associations among linguistic levels. *Lingua* 119(11 The forests behind the trees). 1624–1642. doi:10.1016/j.lingua.2009.02.001.

Stanford, James N. 2012. One size fits all? Dialectometry in a small clan-based indigenous society. *Language Variation and Change* 24(2). 247–278. doi:10.1017/S0954394512000087.

Stoeckle, Philipp. 2014. *Subjektive Dialekträume im alemannischen Dreiländereck*. Hildesheim: Olms. 632 p.

Stoeckle, Philipp. 2016. Horizontal and vertical variation in Swiss German morphosyntax. In Marie-Hélène Côté, Remco Knooihuizen & John Nerbonne (eds.), *The future of dialects: Selected papers from Methods in Dialectology XV (Language Variation 1),* 195–215. Berlin: Language Science Press.

Sui, Daniel Z. 2004. Tobler's first law of geography: A big idea for a small world? *Annals of the Association of American Geographers* 94(2). 269–277. doi:10.1111/j.1467-8306.2004.09402003.x.

Szmrecsanyi, Benedikt. 2012. Geography is overrated. In Sandra Hansen, Christian Schwarz, Philipp Stoeckle & Tobias Streck (eds.), *Dialectological and Folk Dialectological Concepts of Space - Current Methods and Perspectives in Sociolinguistic Research on Dialect Change*, 215–231. Berlin, Boston: De Gruyter.

Szmrecsanyi, Benedikt. 2014. Methods and objectives in contemporary dialectology. In Ilja A. Seržant & Björn Wiemer (eds.), *Contemporary approaches to dialectology: The area of North, Northwest Russian and Belarusian vernaculars / Современные методы в диалектологии. Ареал северных, северо-западных русских и белорусских говоров*, 81–92. Bergen: Department of Foreign Languages, University of Bergen.

Tobler, Waldo R. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46(2). 234–240.

Trudgill, Peter. 1974. Linguistic change and diffusion : Description and explanation in sociolinguistic dialect geography. *Language in Society* 2. 215–246.

Wang, William S.-Y. & Luca L. Cavalli-Sforza. 1986. Spatial distance and lexical replacement. *Language* 62. 38–55.

Warner, Rebecca M. 2013. *Applied Statistics - From Bivariate Through Multivariate Techniques*. 2nd ed. Los Angeles / London / New Delhi / Singapore / Washington D.C.: SAGE.